Published in final edited form as:

Stat Biosci. 2013 May 1; 5(1): 100–118. doi:10.1007/s12561-012-9069-2.

Simultaneous isoform discovery and quantification from RNA-seq

David Hiller and

Center for Epigenetics, Johns Hopkins School of Medicine, 855 N. Wolfe St., Rangos 570, Baltimore, MD 21205 dhiller2@jhmi.edu

Wing Hung Wong

Department of Statistics, Sequoia Hall, 390 Serra Mall, Stanford, CA, 94305
whwong@stanford.edu

Abstract

RNA sequencing is a recent technology which has seen an explosion of methods addressing all levels of analysis, from read mapping to transcript assembly to differential expression modeling. In particular the discovery of isoforms at the transcript assembly stage is a complex problem and current approaches suffer from various limitations. For instance, many approaches use graphs to construct a minimal set of isoforms which covers the observed reads, then perform a separate algorithm to quantify the isoforms, which can result in a loss of power. Current methods also use ad-hoc solutions to deal with the vast number of possible isoforms which can be constructed from a given set of reads. Finally, while the need of taking into account features such as read pairing and sampling rate of reads has been acknowledged, most existing methods do not seamlessly integrate these features as part of the model. We present Montebello, an integrated statistical approach which performs simultaneous isoform discovery and quantification by using a Monte Carlo simulation to find the most likely isoform composition leading to a set of observed reads. We compare Montebello to Cufflinks, a popular isoform discovery approach, on a simulated data set and on 46.3 million brain reads from an Illumina tissue panel. On this data set Montebello appears to offer a modest improvement over Cufflinks when considering discovery and parsimony metrics. In addition Montebello mitigates specific difficulties inherent in the Cufflinks approach. Finally, Montebello can be fine-tuned depending on the type of solution desired.

Keywords

Alternative Splicing; RNA-Seq; Isoform Discovery; Algorithms; Monte Carlo

1 Introduction

Alternative splicing of RNA transcripts is a common phenomenon, with at least 90 – 95 % of human genes known to undergo some form of alternative splicing [36, 27]. Alternative splicing greatly enriches the proteome of a cell by allowing the same gene to produce multiple related protein variants. By differential regulation of splicing, a cell can change the relative concentration of different protein species given different cell types or conditions, and even inactivate a gene by producing a defective version of protein. Alternative splicing can also be important even if it does not directly affect protein production. For instance, untranslated RNAs can play important roles in cells [17], and untranslated regions of RNAs could potentially affect the splicing process. Collectively, all transcript variants of a locus will be termed isoforms even if not translated, and the assembly process of loci with multiple isoforms will be termed isoform discovery.

RNA sequencing is a recent technology that makes possible the integration of two different tasks, isoform discovery and quantification, that previously could only be studied separately. Quantification of splicing events and isoforms could be done with exon or splice junction microarrays [37, 39, 1] but due to the limitations of the probe paradigm, novel splicing events could not be discovered. At the same time, novel isoforms could be discovered by assembling EST reads [18, 39] on top of a splice graph [8], but with limited quantification due to the low count of EST reads.

The past few years have seen an explosion of methods for RNA-seq. Splice-sensitive read mapping is an important first step which we do not consider here other than to note that many methods exist [34, 2, 5]. Isoform quantification alone is a simpler problem than isoform discovery and has been extensively studied using various approaches [26, 14, 15, 32, 19, 16, 20, 12]. Several approaches for identifying differential isoform expression have also been presented [29, 7, 40, 33, 31]. Discovery of novel isoforms is still an important problem [25] since novel junctions are continuing to be discovered even in recent experiments [5]. One class of algorithms for isoform discovery does not align the reads to a reference genome, but directly assembles them by identifying overlapping subsequences of the reads [4, 25]. While innovative and very useful for unsequenced or poorly sequenced organisms, this class of algorithms may have inferior performance when compared to reference-based isoform discovery algorithms when the reference genome is known [25]. The other main class of algorithms aligns reads to a reference genome first, then assembles the mapped read locations to discover the isoforms and will be the focus of this paper. Algorithms in the latter category can be further subdivided into graph-based algorithms and Lasso-based algorithms.

Graph based approaches including Cufflinks [35] and Scripture [6] use graphs to discover the isoforms and are reminiscent of earlier approaches for assembling EST reads, with an adaptation to incorporate information from paired-end reads, and a subsequent step aimed at quantification. More recently Cufflinks has acquired a feature to model the sampling rate [28]. There are some limitations to these approaches. First, Cufflinks is restricted to assembling contiguous overlapping reads, and so may break a low expressed transcript into pieces if there are unobserved junctions or isolated clusters of reads coming from unspliced RNA. Second, graph-based approaches do not adequately integrate the information contained in the read density when discovering the isoforms. Thus they may have difficulty detecting sudden changes in read density which could be due to alternate transcript start or end sites or regions only present in unspliced RNA, as will be shown later. Third, graph approaches do not have a probabilistic way of modeling false positive reads, and instead either incorporate all reads into the model or use a heuristic to filter the reads beforehand. Fourth, statistical power may be lost by breaking up the discovery and quantification steps.

Lasso-based approaches including NSMap [38], IsoLasso [24] and SLIDE [20] construct a path of solutions with increasing numbers of isoforms, where the set of isoforms in each solution is a subset of the set of isoforms in higher solutions, then use model selection criteria to pick the best solution from the path. While these approaches are likelihood-based and thus have a more statistical flavor, the Lasso paradigm is still limited in its generality. For instance, NSMap states that the "right" penalized cost function cannot be optimized under the Lasso model. Also NSMap and IsoLasso do not model the insert length distribution of paired-end reads and IsoLasso only uses a heuristic modification to account for sampling rate while NSMap uses none at all. SLIDE is the most recent and sophisticated of these approaches, but still is limited by the penalty applied to the cost function as well as by the use of GC content to estimate the sampling bias, an approach shown to be not ideal in [23]. There is certainly still a need for a more integrated, comprehensive and tunable approach to RNA sequencing analysis.

2 Methods

We have developed Montebello, a method which addresses many of the statistical issues in current isoform discovery approaches. Montebello assumes reads have been previously mapped to the genome using one of several available read mappers [2, 34]. Montebello implements four steps: three preprocessing steps and a core Monte Carlo algorithm (Figure 1). We have presented an earlier version of this approach in [9].

The first step of Montebello is to determine likely gene loci by scanning the genome for clusters of genomic, junction and paired-end reads. Most existing tools simply use clusters of overlapping reads to define the loci. However, this can lead to difficulties for loci which are expressed at very low read density. Montebello's solution is to divide the genome into bins of fixed width and combine the bins to estimate gene loci. Reads that are less than a bin width apart will be considered part of the same locus. More specifically, bins are joined together if there is a genomic, junction or pair-end read which overlaps both bins. Adjacent bins with average read density higher than a certain threshold are also combined. We only keep those bin clusters with read density exceeding a certain threshold and which contain at least one junction read. While the bin size is arbitrary, the final isoforms found by Montebello will be trimmed so they do not extend beyond the first and last observed read in the region. The default bin size of 1000 can be adjusted.

The second step of Montebello is to define critical points and regions. This greatly simplifies isoform discovery as it restricts the search to those isoforms which are statistically probable. Critical points can be of two types: either splice points supported by at least one junction read, or putative transcription start or end points. A putative transcription start or end point is identified by a statistically significant twofold change in read density. Statistical significance is assessed by assigning a Gamma prior with mean one to the ratio of expression before and after each point in the locus, and computing the posterior mean of the ratio based on the reads in a window of fixed width about the potential point. We consider points which are extrema of the smoothed posterior mean statistic over the length of the locus. Points where the read density increases and which are not within 20 bp after an observed splice end are considered putative transcription start points, while points where the read density decreases and which are not within 20 bp before an observed splice start are considered putative transcription end points. The critical regions are the intervals defined by the critical points.

The third step of Montebello is to categorize the reads into read classes [10, 9, 13, 30, 22] based on the critical regions as follows. For each single- or pair-end read r we consider the set S_r of critical regions overlapped by the read and the set S'_r of critical regions skipped by a splice observed in the read. Then for a set α of included regions and α' of skipped regions, the read class $R_{\alpha,\alpha'}$ consists of all reads r such that $S_r = \alpha$ and $S'_r = \alpha'$. Empty read classes are discarded.

We are finally ready to begin the fourth step, setting up a Monte Carlo-based parallel tempering strategy. In order to do this, for each read class j we compute a sampling rate φ_j . Here we must assume all the reads have length k . Montebello provides two options: assuming a uniform sampling rate for each k -mer, or using a model developed by Li and colleagues [23] which uses the sequence of each k -mer to estimate its sampling rate. Once we have a sampling rate for each k -mer, the sampling rate for a paired-end read is the product of the sampling rates for each constituent single-end read. Then, the sampling rate φ_j for read class j is the sum of the sampling rates for each k -mer r that could be assigned to class j . Since single-end and paired-end reads may both be present in the same sample, and since the sampling rates for these two types of reads are not directly comparable, we must

make an adjustment so that the fraction of the sampling rate which comes from paired-end reads matches the observed proportion of paired-end reads.

We model the observed read count y_j for each read class j as a $\text{Poisson}(\hat{y}_j)$ random variable where:

$$\hat{y}_j = \varphi_j \sum_k \omega_k \delta_{jk} \quad (1)$$

Here ω_k is the expression of isoform k , which is to be estimated by Montebello, and φ_j is the precomputed sampling rate of read class j . The estimated multiplier δ_{jk} adjusts the sampling rate based on the likelihood of observing a read of class j from isoform k as follows. If reads from class j could not have come from isoform k , $\delta_{jk} = 0$. If single-end reads from class j could have come from isoform k , $\delta_{jk} = 1$. If paired-end reads from class j could have come from isoform k , δ_{jk} is proportional to the average over all reads r in class j of the insert length density function evaluated at the insert length of r if r came from isoform k [30]. In this case δ_{jk} can be slightly higher than one if the predicted insert lengths are close to the center of the insert length distribution, or very close to zero if the predicted insert lengths are in the tail of the insert length distribution. For our dataset we assume the insert length density is $N(\mu, \sigma^2)$ with $\mu = 200$, and $\sigma = 20$.

With the above relation in mind, we propose a penalized likelihood which will optimize both the set S of isoforms and their quantities Ω_S at the same time:

$$C(\omega, \varphi, \delta, y) = \sum_j l(y_j, \hat{y}_j) - \lambda K \sum_j y_j \quad (2)$$

Where l is the Poisson log likelihood of the state, K is the total number of isoforms in the model, and λ is a tuning parameter. The summation of the y_j s in the second term has been found to make the penalty more consistent across low and high expressed genes. The L_0 penalty is optimal for producing very sparse solutions, which is necessary because of the high dimensionality of the problem. For instance, a total of 2^n potential isoforms can be described from a locus containing n exon skipping events. However, the L_0 penalty is also nonconvex, which limits Lasso-based approaches such as [38] to using an L_1 or $L_{0.5}$ penalty instead.

One simplification we can make is to note that the function (2) is concave given that the set S of isoforms with nonzero expression is known [14]. Thus we need only optimize over possible sets S , where the likelihood for S is evaluated at the MLE for Ω_S , which is estimated using a coordinatewise Newton's method.

We use parallel tempering [3] to search the space of sets S . Briefly we run ten different Metropolis Hastings simulations at ten different temperatures, allowing the simulations to switch states probabilistically so that more favorable states will be moved to lower temperature simulations where they will be more likely to be retained and improved. Meanwhile high temperature simulations will continue to search the state space for other promising solutions. The temperatures for the simulations are selected in a logarithmic series covering likely temperature jumps when moving from one state to a nearby state, except that the lowest chain has temperature nearly zero so that it will always save the best state encountered so far.

We start each chain with one single-exon isoform and, on each iteration and for each simulation we will choose a proposal state similar to the current state of that simulation using a strategy randomly chosen out of the following six strategies:

1. Removing an isoform
2. Modifying an isoform by adding or removing a splice or changing the transcript start or end site
3. Adding a new isoform which is modified from an existing isoform as in 2
4. Swapping a short portion of two isoforms
5. Swapping a portion of two isoforms and deleting one of them
6. Adding an isoform and swapping a portion of it with another isoform

For each strategy, we choose which isoform(s) to change and which part of the isoform(s) to modify randomly and uniformly. The relative frequency with which Montebello uses each strategy can be tuned.

A further feature we have implemented to improve the results is to emphasize the junctions. For each read class j we compute a factor $s_j = \alpha^{S_j}$ where S_j is the number of splices in read class j (up to 2 for paired-end reads if both ends are junction reads), and α is the junction emphasis parameter. Then in equations 1 and 2 we replace y_j by $y_j^* = y_j s_j$ and ϕ_j by $\phi_j^* = \phi_j s_j$. As we will show below, a reasonable choice is $\alpha = 10$. Emphasizing the junctions discourages Montebello from being especially sensitive to random variability in read density and encourages Montebello to primarily use the splices to build the isoform set, while still using read density to estimate the isoform quantities.

Because of varying gene sizes and structure complexities, we have chosen to adaptively stop Montebello rather than always running it for a fixed number of iterations. We let g be a log likelihood change considered to be small, for instance, 5 % of the total number of read counts (this choice works because the positive term of the Poisson log likelihood is proportional to the read count y_j of class j). We stop the simulation at iteration i if $\delta(i) = l_i - l_{(i-P(i))} < g$, or when the likelihood flattens out over an interval of $P(i)$. As $P(i)$ becomes closer to i Montebello will run longer. For the below results we use $P(i) = 10 \sqrt{i}$, with a maximum of 1000 iterations, and the simulation ran for an average of 650 iterations before stopping. For an example of how Montebello works on a real locus see Fig 2. We have made Montebello available in an R package (<http://www.stanford.edu/group/wonglab/Montebello/Montebello 0.8.tar.gz>)

3 Results

We first confirmed that Montebello is able to give comparable results to Cufflinks on a small simulated data set. We simulated 3.7×10^5 reads from transcript structures of 52 gene loci on chromosome 1 from the KnownGene database [11]. The reads were pair-end with a $N(\mu = 200, \sigma = 100)$ insert length, sampled uniformly with respect to sequence length. We ran Cufflinks for various values of the expression cutoff parameter F (where isoforms with less than $F\%$ of total expression for a particular Cufflinks locus are not reported), and the background expression parameter j (where $j\%$ of the reads are assumed to be from unspliced RNA and discarded). We ran Montebello for a range of values of the tuning parameter λ . Out of 538 annotated junctions used to generate this data set, Montebello could recover 509, while Cufflinks could recover 522. The accuracy of the quantification from the two methods was measured as the Euclidean distance between the normalized expression profile of the estimated and true solutions. Montebello's mean accuracy was also not far behind Cufflinks, with a shortest distance of 0.457 to Cufflinks 0.426. The best accuracies were 0.446 and 0.400, respectively. One advantage of Montebello that we observed is that

Montebello solutions are more parsimonious in the sense that more junctions can be explained using a smaller number of isoforms. (Figure 3)

We next ran Montebello and Cufflinks in parallel on 46.3 million 50 bp paired-end brain reads from an Illumina tissue panel which are available at <http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE19166>. The reads were mapped to the genome (HG18) using SpliceMap 3.36 [2] with 2 mismatches per read allowed. We compared the two methods across 2110 loci discovered by Montebello and in which Cufflinks also identified isoforms. The regions overlapped with 12516 annotated isoforms from the KnownGene database [11], and had lengths ranging from 2 to 504 kb with a median length of 29kb and a median of 14 splices per locus. Montebello discovered a median of 10 isoforms per locus, identifying 38 critical regions, and the median entropy of the isoform distribution per locus was 2.172 ($\lambda = 0.001$). Cufflinks discovered a median of 6 isoforms per locus with a median entropy of 1.829 ($F = 0, j = 0.05$). Cufflinks was selected for comparison with Montebello because it is one of the most comprehensive packages available, as evidenced by its high use and frequent updates.

We show two examples comparing Cufflinks and Montebello on specific loci. In the first example, for gene COMMD4, Cufflinks sometimes has trouble detecting proper transcription start or end sites because it is not very sensitive to sudden shifts in read density (Fig 4). In the second example, for gene TOMM40L, Cufflinks demonstrates a tendency to find short gene fragments or orphan exons (Fig 5). Many of these may be due to a cluster of reads that is separated from other reads by at least one read length. Such a cluster will be seen as a separate isoform by Cufflinks, even though it is likely due to a long low expressed region, perhaps from unspliced RNA. In both cases Montebello was able to correct the problem.

Defining appropriate benchmarks for real data is more difficult than for simulated data because the ground truth is unknown. In place of ground truth, we use a reference set of junctions observed in isoforms in the KnownGene database. We can also define reference sets of alternate 3' and 5' splice site events and of exon skipping events which are observed in the KnownGene database. Alternate exon end events are inferred from a set of splices whenever two of the splices share a common end and the second ends are within 200 bp of each other. Exon skipping events are inferred from a triplet of region-skipping and region-including junctions such that the putative skipped exon has a length of less than 300 bp. For each of the three reference sets, the reference discovery rate is defined as follows: out of the features in each reference set that were observed in the raw data, what proportion were also identified by Cufflinks or Montebello? We expect this number to be close to one. We can also define three novel sets which contain features observed in the raw data which are not found in the KnownGene database. Then the novel discovery rate is the fraction of these features that were identified by Cufflinks or Montebello. Since we expect spurious junctions to be observed in the real data due to mapping and other errors, we do not expect the novel discovery rate to be close to one. However, since there are likely to be bona fide junctions and splicing events not recorded in the KnownGene database, we do not expect the novel discovery rate to be close to zero, either. In general, we would like to see solutions with a lower novel discovery rate, particularly for novel junctions, so that we may be less concerned about the potential for false positives. A third benchmark is complexity, defined as the average number of isoforms observed. Other things being equal, we prefer a less complex solution which does not posit the existence of redundant isoforms.

We demonstrate similar performance between Montebello and Cufflinks over a range of parameter values for 244 loci on chromosome 19 for the above benchmarks. In Figure 6 we plot the novel discovery rate against the reference discovery rate for the junctions, alternate

3' and 5' splice sites, and exon skipping events. The performance curves for Cufflinks and Montebello are similar over much of the parameter space in all three plots. Montebello is able to attain a similar maximum reference discovery rate to Cufflinks. In Figure 7 we plot the complexity against the reference discovery rate for the three feature sets. In all three cases Montebello can generate less complex solutions than Cufflinks solutions with the same reference discovery rate. To compare the quantification of the two methods, we have computed the correlation between the observed read counts and the calculated expressions for each platform (MB: 0.725, Cuff: 0.690) and show a scatter plot for each (Figure 8). Finally, we demonstrate reasonable correlation (0.708) between the Montebello and Cufflinks expression, and show a scatterplot between these two as well (Figure 9).

Finally, we show some examples of how the tunability of Montebello can be leveraged to increase its performance. First, the performance of Montebello can be improved by adjusting the stopping parameters to increase the average running time. We can also improve the quality of the solution by increasing the junction emphasis parameter α up to about 10 or 20 (Fig 11). A value of $\alpha = 10$ was used in this manuscript. Finally, we can adjust the behavior of Montebello by changing the tuning parameter λ (Figure 10). A low value of λ gives us high explanatory power of the observed reads but with higher complexity, while a higher value of λ gives us simpler solutions but with lower explanatory power.

4 Discussion

We have presented Montebello, an isoform discovery algorithm which uses a likelihood-based Monte Carlo algorithm to simultaneously find and quantify the set of isoforms in a locus. We have compared Montebello to Cufflinks, a popular state-of-the-art isoform discovery algorithm, and have seen that Montebello gives results which are similar to or better than Cufflinks in a small simulation and on a real data set. For instance, Montebello can achieve similar detection of reference events annotated in the KnownGene database and better detection of novel events. Montebello also can generate more parsimonious solutions than Cufflinks in simulated and real data sets. Finally, Montebello's tunability can be used to its advantage.

The tunability of Montebello does present a difficulty in terms of selecting appropriate values for the parameters. In our experience we have found λ between 0.01 and 0.2 and α between 5 and 10 to work well. Ideal choice of the parameters depends on the type of solution desired. If recall of known features is most important, a lower value of λ is better. However, if recall should be balanced with parsimony, λ may need to be increased. Since data sets differ in characteristics, it may be desirable to run Montebello on a subset of each data set for several parameter combinations, and compute benchmarks as we have done here.

In addition to the novelty of the results, Montebello makes methodological developments to the field of RNA-Seq analysis. Early approaches including Cufflinks [35] and Scripture [6] take on a graph-based, heuristic approach. Later Lasso-based approaches, though having more mathematical rigor, lack the ability to model certain features which Montebello models in a natural way. For instance, neither IsoLasso [24] nor NSMap [38] models the insert length distribution of paired-end reads and IsoLasso only uses a heuristic modification to account for sampling rate while NSMap uses none at all. Montebello models the sampling rate and insert length as part of the core statistical model rather than as a separate step as in Cufflinks and IsoLasso. Montebello can also process single-end and paired-end reads at the same time rather than separately, as in Scripture. While all the Lasso-based methods solve a penalized likelihood, IsoLasso and SLIDE use an L1 penalty and NSMap tries an L1 and L0.5 penalty, while Montebello is able to utilize an L0 penalty which is more ideal for this problem [38]. Finally, Montebello is the only approach thus far to use a Monte Carlo algorithm to

determine the isoform structure. While Montebello is more complex and takes longer to run than other algorithms, the additional complexity may be merited for such a difficult problem as isoform discovery.

One limitation of Montebello is that it currently assumes a fixed read length, and may return less accurate results if reads have variable lengths. Montebello also does not currently infer the strand of isoforms by using the XS SAM tag which gives the inferred strand of junction reads [21]. However, if we know the strand of all reads, we can prepare a separate SAM file for the reads on each strand and run Montebello twice.

A particular difficulty with Montebello is the time required to run it, and in this final section we try to understand why, and give tips on making Montebello more scalable. A raw list of 4617 loci took a total of 21.4 days of computer cluster time, or 6.67 minutes per locus, to run, with each chromosome processed on a separate core. The long running time is due mostly to the nested nature of the algorithm. For a typical run, for each of ten simulations and on each of 600 iterations we performed a transcript update and then a Newton simulation of up to 100 iterations to quantify the isoforms. In this case the inner loop would run up to 600000 times and compute several matrix products each time. Around a third of the running time of Montebello was taken up by preprocessing, including fitting the sampling rate model separately for each read.

It is helpful to note that the time scales in the number of gene loci considered rather than the total number of reads. While some read-dependent setup must be performed on the raw SAM files to determine the loci, this setup took only on the order of hours for 46.3 million reads compared to 21.4 days for the remainder of the algorithm. Thus a much larger number of reads could be processed as a single sample without too much gain in running time.

Once the read-dependent setup has been completed, Montebello is completely parallelizable by gene locus. Montebello has been modularized to separate the read-dependent and read-independent steps to make it easier to run in parallel. Unix scripts to assist in creating and running parallel R scripts are available from the author upon request.

If Montebello is still taking too long to run, many parameters can be changed to reduce the running time, although quite possibly at the expense of quality of the solution. For instance, we can run with fewer than ten parallel simulations, we can change the stopping time parameters, or we could change the default Poisson model to a normal model which allows for direct computation of the MLE for the transcript abundances, although this model is significantly less robust. Since the running time for each locus scales linearly in the number of read classes in that locus, we can also save time by considering each read as an unpaired single-end read. Finally, we can assume a uniform sampling rate rather than modeling the sampling rate based on the sequence.

Finally, Montebello can be run only on a single set of reads at a time. Therefore, to make Montebello scalable for multiple runs of RNA-seq, we recommend the following approach. First combine reads from all runs of RNA-Seq into a single SAM file. Run Montebello once on this file to generate a set of loci and a set of isoforms for each locus. Then for each run separately, run transcript quantification on the reads from each locus using the isoforms identified by the initial run. Downstream analysis can be performed using other software packages, such as tools put out by the makers of Cufflinks.

Acknowledgments

We thank Hui Jiang and Nicholas Johnson for useful discussions. DH developed and tested the model. WHW initiated and supervised the project. DH drafted and WHW revised the paper. DH was funded a Ric Weiland

Graduate Fellowship (Stanford University) and by NIH grants R01 HG004634 and R01 HG005220. WHW was supported by NIH grants R01 HG004634 and R01 HG005717.

References

- Anton MA, Gorostiaga D, Guruceaga E, Segura V, Carmona-Saez P, Pascual-Montano A, Pio R, Montuenga LM, Rubio A. Space: an algorithm to predict and quantify alternatively spliced isoforms using microarrays. *Genome Biol.* 2008; 9:R46. [PubMed: 18312629]
- Au KF, Jiang H, Lin L, Xing Y, Wong WH. Detection of splice junctions from paired-end rna-seq data by splicemap. *Nucl. Acids Res.* 2010; 38(14):4570–4578. [PubMed: 20371516]
- Geyer, C. Markov chain monte carlo maximum likelihood. In: Keramidas, EM., editor. *Computing Science and Statistics: Proc. 23rd Symposium on the Interface*. Fairfax Station, VA: Interface Foundation; 1991. p. 156-163.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. Full-length transcriptome assembly from rna-seq data without a reference genome. *Nat Biotech.* 2011; 29:644–652.
- Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, Stoeckert CJ, Hogenesch JB, Pierce EA. Comparative analysis of rna-seq alignment algorithms and the rna-seq unified mapper (rum). *bioinformatics.* 2011; 27(18):2518–2528. [PubMed: 21775302]
- Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincnas. *Nat Biotech.* 2010; 28:503–510.
- Hardcastle T, Kelly K. bayseq: Empirical bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics.* 2010; 11(1):422. [PubMed: 20698981]
- Heber S, Alekseyev M, Sze SH, Tang H, Pevzner PA. Splicing graphs and EST assembly problem. *Bioinformatics.* 2002; 18(suppl 1):S181–188. [PubMed: 12169546]
- Hiller, D. Ph.D. thesis. Stanford University; 2010. Alternative splicing analysis using rna-seq data.
- Hiller D, Jiang H, Xu W, Wong WH. Identifiability of isoform deconvolution from junction arrays and rna-seq. *Bioinformatics.* 2009; 25(23):3056–3059. [PubMed: 19762346]
- Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. The ucsc known genes. *Bioinformatics.* 2006; 22(9):1036–1046. [PubMed: 16500937]
- Hu M, Zhu Y, Taylor J, Liu J, Qin Z. Using poisson mixed-effects model to quantify transcript-level gene expression in rna-seq. *Bioinformatics.* 2012; 28(1):63–68. [PubMed: 22072384]
- Jiang, H. Ph.D. thesis. Stanford University; 2009. Computational and statistical approaches in rna sequencing analysis.
- Jiang H, Wong W. Statistical inferences for isoform expression in rna-seq. *Bioinformatics.* 2009; 25(8):1026–1032. [PubMed: 19244387]
- Katz Y, Wang ET, Airolidi EM, Burge CB. Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nat Meth.* 2010; 7:1009–1055.
- Kim H, Bi Y, Pal S, Gupta R, Davuluri R. Isoformex: isoform level gene expression estimation using weighted non-negative least squares from mrna-seq data. *BMC Bioinformatics.* 2011; 12(1):305. [PubMed: 21794104]
- Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE. Unproductive splicing of sr genes associated with highly conserved and ultraconserved dna elements. *Nature.* 2007; 446:926–929. [PubMed: 17361132]
- Lee C. Generating consensus sequences from partial order multiple sequence alignment graphs. *Bioinformatics.* 2003; 19(8):999–1008. [PubMed: 12761063]
- Lee S, Seo CH, Lim B, Yang JO, Oh J, Kim M, Lee S, Lee B, Kang C, Lee S. Accurate quantification of transcriptome from rna-seq data by effective length normalization. *Nucleic Acids Research.* 2011; 39(2):e9. [PubMed: 21059678]
- Li B, Dewey C. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics.* 2011; 12(1):323. [PubMed: 21816040]

21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPD. The sequence alignment/map format and samtools. *Bioinformatics*. 2009; 25(16):2078–2079. [PubMed: 19505943]
22. Li J, Jiang C, Brown J, Huang H, Bickel P. Sparse linear modeling of next-generation mrna sequencing (rna-seq) data for isoform discovery and abundance estimation. *Proc Natl Acad Sci*. 2011; 108(50):19,867–19,872.
23. Li J, Jiang H, Wong W. Modeling non-uniformity in short-read rates in rna-seq data. *Genome Biology*. 2010; 11(5):R50. [PubMed: 20459815]
24. Li W, Feng J, Jiang T. Isolasso: A lasso regression approach to rna-seq based transcriptome assembly. *Journal of Computational Biology*. 2011; 18(11):1693–1707. [PubMed: 21951053]
25. Martin JA, Wang Z. Next-generation transcriptome assembly. *Nat Rev Genet*. 2011; 12(10):671–682. [PubMed: 21897427]
26. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods*. 2008; 5(7):621–628. [PubMed: 18516045]
27. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*. 2008; 40:1413–1415. [PubMed: 18978789]
28. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving rna-seq expression estimates by correcting for fragment bias. *Genome Biology*. 2011; 12:R22. [PubMed: 21410973]
29. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010; 26(1):139–140. [PubMed: 19910308]
30. Salzman J, Jiang H, Wong W. Statistical modeling of rna-seq data. *Statistical Sciences*. 2011; 26(1):62–83.
31. Shen S, Won Park J, Huang J, Dittmar K, Lu Z, Zhou Q, Carstens R, Xing Y. Mats: a bayesian framework for flexible detection of differential alternative splicing from rna-seq data. *Nucleic Acids Res*. 2012; 40(8):e61. [PubMed: 22266656]
32. Stegle, O.; Drewe, P.; Bohnert, R.; Borgwardt, K.; Rätsch, G. Statistical tests for detecting differential rna-transcript expression from read counts. 2010. Available on Nature Precedings, <http://precedings.nature.com/documents/4437/version/1>
33. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in rna-seq: A matter of depth. *Genome Research*. 2011 URL <http://genome.cshlp.org/content/early/2011/10/28/gr.124321.111.abstract>.
34. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009; 25(9):1105–1111. [PubMed: 19289445]
35. Trapnell C, Williams BA, Pertea G, Mortazavi AM, Kwan G, van Baren MJ, Salzberg S, Wold B, Pachter L. Transcript assembly and abundance estimation from rna-seq reveals thousands of new transcripts and switching among isoforms. *Nature Biotechnology*. 2010; 28:511–515.
36. Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008; 456:470–476. [PubMed: 18978772]
37. Wang H, Hubbell E, Hu JS, Mei G, Cline M, Lu G, Clark T, Siani-Rose MA, Ares M, Kulp DC, Haussler D. Gene structure-based splice variant deconvolution using a microarray platform. *Bioinformatics*. 2003; 19:i315–i322. [PubMed: 12855476]
38. Xia Z, Wen J, Chang CC, Zhou X. Nsmap: A method for spliced isoforms identification and quantification from rna-seq. *BMC Bioinformatics*. 2011; 12(1):162. URL <http://www.biomedcentral.com/1471-2105/12/162>. [PubMed: 21575225]
39. Xing Y, Yu T, Wu YN, Roy M, Kim J, Lee C. An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucl. Acids Res*. 2006; 34(10):3150–3160. [PubMed: 16757580]
40. Zhou YH, Xia K, Wright FA. A powerful and exible approach to the analysis of rna sequence count data. *Bioinformatics*. 2011; 27(19):2672–2678. [PubMed: 21810900]

Fig. 1. Schematic showing how Montebello works on a toy example. (a) Identify loci: Bin genome into bins of 1000 bp. Find clusters of bins that are known to be part of the same locus because of junction or paired-end reads that overlap multiple bins. (b) Divide locus into critical regions based on observed splice points and inferred transcription start and end points. (c) Classify reads into read classes based on membership in critical regions. Black: overlaps critical region, white: skips critical region, gray: unknown. Reads a - i are classified into read classes A - G. (d) Run core Montebello algorithm to simultaneously find ideal isoform structure (left, black) and quantification (right, gray), and trim resulting isoforms so they extend no further than the furthest observed read in the locus.

Fig. 2.

Example of how MonteBello works on a real locus with $\lambda = 0.001$. (a) The UCSC browser shows three isoforms and two alternative splicing events for the gene COMMD4: an alternative 5' splice site and an exon skipping event. (b) Observed read counts (vertical gray bars) and splices (horizontal black bars, thickness correlates to expression). Neither alternative splicing event seen in the UCSC browser is supported by observed junction reads, but two novel exon skipping events can be inferred as well as an alternate 5' splice site. (c) Iteration 10. Each row represents one isoform. The black track on the left represents the structure of the isoform, and the gray bar on the right represents the relative quantity of the isoform predicted by MonteBello. From an initial state of a single intronless isoform,

Montebello has already discovered several of the splices. (d) Iteration 100. Montebello has discovered one of the two exon skipping events as well as the alternate 5' splice site. Montebello is also modeling some intronic regions in the 5' end of the transcript which may not be desirable. On inspection, however, these regions have higher read density than some of the exons toward the 3' end of the transcript. (e) Iteration 300. Though Montebello can be run an arbitrary length of time, by this time a nearly optimal solution has been produced with both alternative splicing events discovered.

Fig. 3.

Plot of average number of isoforms against number of explained junctions for Montebello solutions at various values of λ and Cufflinks solutions at various values of the parameters F and j . The vertical dotted line on the right is the total number of junctions in the data set. Montebello is better at finding very parsimonious solutions and can explain 480 to 500 reads with significantly fewer isoforms than Cufflinks.

Fig. 4.

Example contrasting the behavior of Cufflinks ($F = 0$, $j = 0.05$) and Montebello ($\lambda = 0.001$) for the COMMD4 locus in Figure 2. (a) and (b) are the same as in Figure 2 and show the UCSC annotation and raw reads for the locus. (c) Isoform structure and quantification predicted by Cufflinks. Cufflinks identified one of the two exon skipping events and the alternate 5' splice site. Cufflinks seemed to have trouble discovering the correct transcription end site because of a few false positive reads. (d) Montebello correctly identified both exon skipping events and the alternate 5' splice site. Montebello generated a less parsimonious solution than Cufflinks here, and the transcription start site may be slightly too early. However, Montebello found the correct transcription end site, which may be due to Montebello's statistical method for identifying changes in read density.

Fig. 5.

Example contrasting the behavior of Cufflinks and Montebello on a second gene locus of gene TOMM40L. (a) UCSC shows four isoforms of TOMM40L in this locus with one exon skipping event, one alternate transcription start site and one alternate transcription end site. (b) Genomic read and splice counts. We cannot infer the annotated exon skipping event from the observed reads. However, two novel exon skipping junctions can be observed. It is difficult to discern whether the reads support either the alternate transcription start site or the alternate transcription end site. (c) Cufflinks isoforms ($F = 0$, $j = 0.05$). Cufflinks does not discover any transcript that extends the whole length of TOMM40L, but rather discovers a series of transcript fragments. However, both exon skipping events and the alternate transcript start site are correctly identified. (d) Montebello isoforms ($\lambda = 0.001$). The second Montebello isoform closely matches the top annotated isoform of TOMM40L. The top Montebello isoform may be similar to an unannotated isoform with an alternate transcription

end site. Montebello correctly identifies both exon skipping events. However, the transcription start site comes a little bit too early. Also the bottom Montebello isoform may be picking up on the presence of partially spliced RNA.

Fig. 6.

(a) Plots the novel discovery rate against the reference discovery rate for junctions, for 244 gene loci from chromosome 19. Compares Montebello solutions with λ ranging from 10^{-5} to 0.2 and Cufflinks solutions with F ranging from 0 to 0.2 and $j = 0.05$ or 0. If we are concerned only with detection of reference junctions, the plot functions as an ROC curve, and Montebello performs similarly to Cufflinks. However, since many of the novel junctions may be real, it is possible that the ability of Montebello to obtain a higher novel discovery rate is an advantage. (b) Plots the novel discovery rate against the reference discovery rate for alternate 3' and 5' splice site events. Again, the two curves are similar except that Montebello can attain a higher novel discovery rate. We have somewhat more confidence that these novel events are real because they involve two observed splices and one matching end. (c) Plots the novel discovery rate against the reference discovery rate for exon skipping events. Again, the two curves are similar, and Montebello can still attain a slightly higher novel discovery rate. We are the most confident that these novel events are real because they involve three observed splices and two matching ends.

Fig. 7.

(a) Plots the average number of isoforms discovered per locus against the reference discovery rate for junctions, for 244 gene loci from chromosome 19. Compares Montebello solutions with λ ranging from 10^{-5} to 0.2 and Cufflinks solutions with F ranging from 0 to 0.2 and $j = 0.05$ or 0. While both methods can detect approximately the same number of annotated splices, Montebello is considerably more parsimonious in that it can generate less complex solutions with the same explanatory power as Cufflinks. (b) Plots the average number of isoforms discovered per locus against the reference discovery rate for alternative 3' and 5' splice sites. Again, Montebello is more parsimonious. (c) Plots the average number of isoforms discovered per locus against the reference discovery rate for exon skipping events. Montebello is more parsimonious than Cufflinks for discovery rates up to 80% and similar to Cufflinks for discovery rates above 80%.

Fig. 8.

(a) Plot of Cufflinks log expression vs log number of observations for each splice. (b) Plot of Montebello log expression vs log number of observations for each splice.

Fig. 9.
Plot of Montebello log expression vs Cufflinks log expression.

Fig. 10.

Example showing the effect of changing the tuning parameter λ for the locus of Fig 2. (a) and (b) show the UCSC annotation and observed reads in the COMMD4 locus as before. (c) $\lambda = 1$: 1 isoform. For high values of λ Montebello usually returns only the dominant isoform since the cost of adding additional isoforms is so high. (d) $\lambda = 0.2$: 2 isoforms. One of the splicing events is not discovered here, and two are combined. (e) $\lambda = 0.01$: 5 isoforms. Both exon skipping events and the alternate 3' splice site have been discovered, and a fifth isoform models unspliced RNA. (d) and (e) illustrate that for intermediate values of λ , Montebello returns solutions that balance parsimony and explanatory power. (f) $\lambda = 2$

* 10^{-5} : 10 isoforms. For low values of λ Montebello fits most of the observed splices, but also tends to over fit the read density, resulting in many redundant isoforms.

Fig. 11.

(a) For $\lambda = 0.001$ and using the 244 reads from chromosome 19, shows the effect of varying the junction emphasis parameter α from 1 to 100, by plotting the number of isoforms against the reference discovery rate for junctions. The plot shows that a higher α produces a simpler solution with highest reference discovery rate up to $\alpha = 20$. (b) Shows the effect of varying α from 1 to 100 by plotting the novel discovery rate against the reference discovery rate for junctions. Although $\alpha = 20$ has a higher reference discovery rate, we are careful in setting α too high because the solution will be heavily influenced by only a few read classes, so we use $\alpha = 10$.