1 Overview and motivation

This document aims to improve upon our current RSEM_eval model – which is already quite good – in two ways. First, we explicitly model both transcripts and contigs. In the current model, we model only the contigs explicitly, not the transcripts, and this seems to me to have led to some difficulties in specifying the model distribution. Second, we find (under minimal assumptions) the exact distribution of the contigs, given the transcripts and reads. In the current model, only an approximate distribution is used. In fact, as far as I can tell, this document is the first place where the exact distribution of the contigs from a transcript (or a chromosome, for that matter) has been given, at least under such realistic assumptions.

In Section 2, we define random variables and fix notation. In Section 3, we find/stipulate the joint distribution of the random variables under certain assumptions. In Section 4, we describe a simple heuristic approach to find the mode of the joint distribution. Finally, in Section **??**, we describe how to approximately integrate over latent variables in order to get our score.

2 Notation

First, we define a contig. A contig is defined to be a maximal contiguous subsequence of a transcript such that every base is covered by at least one read. This corresponds to the definition used by `collectContigs` with window size 0.

We define the following constants.

1. numReads is the total number of reads.
2. readLen is the length of each read. We assume that all reads have the same length.

We define the following random quantities:

1. NumTrans is the number of distinct transcripts (in the underlying sequenced sample).
2. TranLen_i is the length of transcript i.
3. TranSeq_i is the sequence of transcript i.
4. Expr_i is number of reads that (truly) came from transcript i.
5. Read_n is the sequence of read n.
6. NumContigs is the number of distinct contigs.
7. ContigLen_j is the length of contig j.
8. ContigSeq_j is the sequence of transcript j.
9. Coverage_j is the number of reads that (truly) came from contig j.
10. $\text{NumContigsFromTran}_i$ is number of contigs which originated from transcript i.
11. ContigsFromTran_i is the set of indices of the contigs which originated from transcript i.

For our convenience, we group the above as follows:

1. $\text{Tran}_i = (\text{TranLen}_i, \text{TranSeq}_i)$.
2. $\text{Tran} = (\text{Tran}_i : i = 1, \ldots, \text{NumTrans})$.
3. $\text{Expr} = (\text{Expr}_i : i = 1, \ldots, \text{NumTrans})$.
4. $\text{Trans} = (\text{NumTrans}, \text{Tran})$.

5. $\text{Contig}_j = (\text{ContigLen}_j, \text{ContigSeq}_j)$.
6. $\text{Contig} = (\text{Contig}_j : j = 1, \ldots, \text{NumContigs})$.
7. $\text{Coverage} = (\text{Coverage}_j : j = 1, \ldots, \text{NumContigs})$.
8. $\text{Contigs} = (\text{NumContigs}, \text{Contig})$.
9. $\text{Contig}_{(i)} = (\text{Contig}_j : j \in \text{ContigsFromTran}_i)$.
10. $\text{ContigsFromTrans} = (\text{NumContigsFromTran}_i, \text{ContigsFromTran}_i : i = 1, \ldots, \text{NumTrans})$.
11. $\text{Reads} = (\text{Read}_n : n = 1, \ldots, \text{numReads})$.

Sequence indices start at 0. Other indices start at 1.

Unless there is ambiguity, for random variables X and Y, we will write $P(x)$ instead of $P(X = x)$, and $P(x|y)$ instead of $P(X = x|Y = y)$. We write $X =_D Y$ if X equals Y in distribution. All indices within sequences start at 0.

Our score is $\text{score}(\text{reads}, \text{contigs}) = P(\text{reads}, \text{contigs})$. Of course, we are really interested in the normalized version $P(\text{contigs}|\text{reads})$, but it is not necessary to normalize when we are comparing assmeblies based on a fixed set of reads. The score is related to the joint distribution of all random variables listed above by marginalization:

$$\begin{aligned}&\text{score}(\text{reads}, \text{contigs}) = P(\text{reads}, \text{contigs})\\ &= \int P(\text{reads}, \text{contigs}, \text{coverage}, \text{trans}, \text{expr}, \text{contigsFromTrans})\, d\text{coverage}\, d\text{trans}\, d\text{expr}\, d\text{contigsFromTrans}\end{aligned}$$

We want to evaluate the score, and we will do this by a Laplace-like approximation, as follows: (i) specify the joint distribution, (ii) find a mode or near-mode of of this distribution, (iii) perform a mode-based approximation of the integral over hidden variables.

3 The joint distribution

We decompose the joint distribution as follows, using the chain rule:

$$\begin{aligned}P(\text{reads}, \text{trans}, \text{expr}, \text{contigs}, \text{coverage}, \text{contigsFromTrans}) = {} & P(\text{trans}, \text{expr})\\ & \cdot P(\text{reads}|\text{trans}, \text{expr})\\ & \cdot P(\text{contigs}, \text{coverage}, \text{contigsFromTrans}|\text{reads}, \text{trans}, \text{expr})\end{aligned}$$

In the rest of this section, we will stipulate functional forms for each of these distributions.

3.1 Distribution of $\text{Trans}, \text{Expr}$

For the prior distribution $P(\text{trans}, \text{expr})$, we stipulate, similarly to the current RSEM_eval model, that:

1. $\text{NumTrans} \sim \text{Poisson}(\mu_0)$.
2. $\text{TranLen}_i|\{\text{NumTrans} = \text{numTrans}\} \sim \text{NegativeBinomial}(r_0, p_0)$, for $i = 1, \ldots, \text{numTrans}$.
3. $\text{TranSeq}_i|\{\text{NumTrans} = \text{numTrans}, \text{TranLen}_i = \text{tranLen}_i\} \sim \text{Uniform}(\{\texttt{A}, \texttt{T}, \texttt{G}, \texttt{C}\}^{\text{tranLen}_i})$.
4. $\text{Expr}_i|\{\text{NumTrans} = \text{numTrans}, \text{TranLen}_i = \text{tranLen}_i, \text{TranSeq}_i = \text{tran}_i\} \sim \text{Uniform}(\{1, 2, \ldots, \text{numReads}\})$. (A non-uniform distribution for Expr_i might be more reasonable.)

So

$$
\begin{aligned}
&P(\text{trans}, \text{expr}) \\
&= P(\text{numTrans}) \prod_{i=1}^{\text{numTrans}} P(\text{tranLen}_i|\text{numTrans}) \cdot P(\text{tranSeq}_i|\text{numTrans}, \text{tranLen}_i) \cdot P(\text{expr}_i|\text{numTrans}, \text{tranLen}_i, \text{tranSeq}_i) \\
&= \text{Poisson}(\text{numTrans}|\mu_0) \prod_{i=1}^{\text{numTrans}} \text{NegativeBinomial}(\text{tranLen}_i|r_0, p_0) \cdot 4^{-\text{tranLen}_i} \cdot \text{numReads}^{-1} \\
&= \frac{\mu_0^{\text{numTrans}}}{\text{numTrans}!} e^{-\mu_0} \prod_{i=1}^{\text{numTrans}} \binom{\text{tranLen}_i + r_0 - 1}{\text{tranLen}_i} (1-p_0)^{r_0} p_0^{\text{tranLen}_i} \cdot 4^{-\text{tranLen}_i} \cdot \text{numReads}^{-1} \\
&= \frac{(\mu_0(1-p_0)^{r_0}/\text{numReads})^{\text{numTrans}}}{\text{numTrans}!} e^{-\mu_0} \prod_{i=1}^{\text{numTrans}} \binom{\text{tranLen}_i + r_0 - 1}{\text{tranLen}_i} (p_0/4)^{\text{tranLen}_i}
\end{aligned}
$$

where (μ_0, r_0, p_0) are the prior parameters.

3.2 Distribution of $\text{Reads}|\{\text{Trans}, \text{Expr}\}$

For the read likelihood, we use RSEM:

$$P(\text{reads}|\text{trans}, \text{expr}) = P_{RSEM}(\text{reads}|\text{trans}, \text{expr})$$

No correction is needed because we are conditioning on the actual transcripts, not contigs.

3.3 Distribution of $(\text{Contigs}, \text{Coverage}, \text{ContigsFromTrans})|\{\text{Reads}, \text{Trans}, \text{Expr}\}$

3.3.1 Assumptions

We make the following assumptions/stipulations.

Assumption 1. First, just as in the current RSEM_eval approach, we assume that the contigs and their coverage are conditionally independent of the particular read sequences, given the transcripts and their expression. In symbols:

$$
\begin{aligned}
&P(\text{contigs}, \text{coverage}, \text{contigsFromTrans}|\text{reads}, \text{trans}, \text{expr}) \\
&= P(\text{contigs}, \text{coverage}, \text{contigsFromTrans}|\text{trans}, \text{expr})
\end{aligned}
$$

Recall that the numReads and readLen are not random variables; indeed, the contig and coverage distribution depends crucially on numReads and readLen.

The practical consequence of this first stipulation is that we will (a) determine (exactly) the probability that (i) a certain number of contigs, with (ii) certain lengths and (iii) certain coverages, come from a transcript with a particular length and expression, and (b) we will apply this general distribution to our particular case.

Assumption 2. Second, we stipulate that different assignments of contigs as coming from particular transcripts are equivalent, provided that each transcript has equal numbers of contigs with particular lengths and coverages under all the different such assignments. In symbols, let $\text{ContigCounts}_i = (\text{ContigCounts}_{i,l,c})$ and $\text{ContigCounts}_{i,l,c} = |\{j \in \text{ContigsFromTran}_i : \text{ContigLen}_j = l, \text{Coverage}_j = c\}|$. Define the following equivalence relation:

$$\text{ContigsFromTrans} \sim \text{ContigsFromTrans}' \iff (\text{ContigCounts}_{i,l,c} = \text{ContigCounts}'_{i,l,c} \text{ for all } i = 1, \ldots, \text{NumTrans})$$

Our stipulation is that if $\text{ContigsFromTrans} \sim \text{ContigsFromTrans}'$ holds, then the following events are identical (for any set A):

$$\{\omega : \text{ContigsFromTrans}(\omega) \in A\} = \{\omega : \text{ContigsFromTrans}'(\omega) \in A\}$$

The practical consequence of this second stipulation is that we will not need to normalize by the total number of ways to assign contig indices to transcripts. It would be technically easy to do so, but it does not seem like a good idea, because the labelling of the contigs is not meaningful.

Assumption 3. Third, we stipulate that the start positions of the reads coming from each transcript are (i) independent and (ii) uniformly distributed among the possible start positions. The practical consequence of this third assumption is that the probability question will reduce to a combinatorial question.

Assumption 4. Fourth, we stipulate that the contigs (and their coverage) from a particular transcript are conditionally independent (given that transcript and its expression) of other transcripts, their expression, their contigs, and their contigs' coverages. In symbols:

$$\begin{aligned}&P(\text{contigs}, \text{coverage}, \text{contigsFromTrans} | \text{trans}, \text{expr})\\ &= \prod_{i=1}^{\text{numTrans}} P(\text{contigs}_{(i)}, \text{coverage}_{(i)}, \text{contigsFromTran}_i | \text{tran}_i, \text{expr}_i)\end{aligned}$$

This assumption seems to be more or less true.

3.3.2 Distribution

Theorem 1. *Under Assumptions 1–3,*

$$\begin{aligned}&P(\text{contig}_i, \text{coverage}_i, \text{contigsFromTran}_i | \text{tran}_i, \text{expr}_i)\\ &= \frac{\binom{\text{numContigsFromTran}_i}{\text{contigCounts}_i}\binom{\text{tranLen}_i+1-\sum_{j\in\text{contigsFromTran}_i}\text{contigLen}_j}{\text{numContigsFromTran}_i}\prod_{j\in\text{contigsFromTran}_i}\binom{\text{coverage}_j-1}{\text{contigLen}_j-\text{readLen}}_{\text{readLen}+1}}{\binom{\text{expr}_i+\text{tranLen}_i-\text{readLen}}{\text{expr}_i}}\end{aligned}$$

where $\binom{n}{\mathbf{k}}$ denotes a multinomial coefficient, $\binom{n}{k}$ denotes a binomial coefficient, and $\binom{n}{k}_N$ denotes a polynomial coefficient, defined below.

Corollary. Due to Assumption 4,

$$\begin{aligned}&P(\text{contigs}, \text{coverage}, \text{contigsFromTran}_i | \text{trans}, \text{expr})\\ &= \prod_{i=1}^{\text{numTrans}} P(\text{contig}_i, \text{coverage}_i, \text{contigsFromTran}_i | \text{tran}_i, \text{expr}_i)\\ &= \prod_{i=1}^{\text{numTrans}} \frac{\binom{\text{numContigsFromTran}_i}{\text{contigCounts}_i}\binom{\text{tranLen}_i+1-\sum_{j\in\text{contigsFromTran}_i}\text{contigLen}_j}{\text{numContigsFromTran}_i}\prod_{j\in\text{contigsFromTran}_i}\binom{\text{coverage}_j-1}{\text{contigLen}_j-\text{readLen}}_{\text{readLen}+1}}{\binom{\text{expr}_i+\text{tranLen}_i-\text{readLen}}{\text{expr}_i}}\end{aligned}$$

3.3.3 Polynomial coefficients

Before proving Theorem 1, we will define polynomial coefficients, look at their combinatorial interpretation, and list some of their properties. Most of this information is from Fahssi (2012) and Balasubramanian et al. (1995).

The polynomial coefficient $\binom{n}{k}_N$ may be defined as follows:

$$\binom{n}{k}_N = \begin{cases} 0 & \text{if } n < 0 \text{ or } k < 0 \text{ or } k > (N-1)\cdot n \\ 1 & \text{if } n = 0 \text{ and } k = 0 \\ \sum_{j=0}^{N-1} \binom{n-1}{k-j}_N & \text{otherwise} \end{cases}$$

For computation, the following equivalent formula is more useful (Balasubramanian et al. (1995), Eq. (7), citing other work):

$$\binom{n}{k}_N = \begin{cases} 0 & \text{if } n < 0 \text{ or } k < 0 \text{ or } k > (N-1)\cdot n \\ 1 & \text{if } n = 0 \text{ and } k = 0 \\ \sum_{a=0}^{\min(n,\lfloor k/N \rfloor)} (-1)^a \binom{n}{a} \binom{k+n-a\cdot N-1}{n-1} & \text{otherwise} \end{cases}$$

The name "polynomial coefficient" comes from the fact that $\binom{n}{k}_N$ is the coefficient of t^k in the polynomial $(1+t+t^2+\cdots+t^{N-1})^n$. Thus the polynomial coefficient is a generalization of the binomial coefficient, in the sense that $\binom{n}{k}_2 = \binom{n}{k}$.

The combinatorial interpretation of the polynomial coefficients is as follows. The number of distinct ways in which k (unlabelled) balls can be placed in n (labelled) urns, allowing at most $N-1$ balls to fall in each urn, is $\binom{n}{k}_N$. For example, the number of ways to place $k = 4$ balls in $n = 3$ urns, allowing at most $N-1 = 3-1 = 2$ balls to fall in each urn, is 6, namely:

1. Urn 1: 2 balls, Urn 2: 2 balls, Urn 3: 0 balls.
2. Urn 1: 2 balls, Urn 2: 0 balls, Urn 3: 2 balls.
3. Urn 1: 0 balls, Urn 2: 2 balls, Urn 3: 2 balls.
4. Urn 1: 2 balls, Urn 2: 1 ball, Urn 3, 1 ball.
5. Urn 1: 1 ball, Urn 2: 2 balls, Urn 3, 1 ball.
6. Urn 1: 1 ball, Urn 2: 1 ball, Urn 3, 2 balls.

Indeed, $\sum_{a=0}^{\min(3,\lfloor 4/3 \rfloor)} (-1)^a \binom{3}{a} \binom{4+3-a\cdot 3-1}{3-1} = (-1)^0 \binom{3}{0} \binom{6-3\cdot 0}{2} + (-1)^1 \binom{3}{1} \binom{6-3\cdot 1}{2} = 6$.

3.3.4 Read placements

Our proof of Theorem 1 is based on counting "read placements". To define a read placement of expr_i reads of length readLen within a transcript of length tranLen_i, fix an arbitrary enumeration $1, 2, \ldots, \text{expr}_i$, of the reads. A read placement is a choice of start positions $s = (s_1, s_2, \ldots, s_{\text{expr}_i})$, one for each read, which is required to satisfy: (i) $s_1 \geq 0$, (ii) $s_{k-1} \leq s_k \leq s_{k+1}$, and (iii) $s_{\text{expr}_i} \leq \text{tranLen}_i - \text{readLen}$.

By fixing an arbitrary enumeration in advance, we are stipulating that the ordering of reads within the placement does not matter. Since all reads have the same length, there are $\text{tranLen}_i - \text{readLen} + 1$ possibile read start positions. Thus each read placement is equivalent to a choice of expr_i unordered read start positions, with replacement, from $\text{tranLen}_i - \text{readLen} + 1$ possibilities. This gives the following lemma:

Lemma 2. *The total number of read placements, as defined above, is* $\binom{\text{expr}_i + \text{tranLen}_i - \text{readLen}}{\text{expr}_i}$.

Proof. In general, the number of ways to choose k unordered elements, with replacement, from a collection of size n, is $\binom{n+k-1}{k-1} = \binom{n+k-1}{n}$. The lemma follows from our previous remarks. ■

We are most interested in read placements that result in contigs of particular sizes. To start, it is useful to know how many read placements fully cover the transcript, i.e., how many read placements result in exactly one contig that equals the transcript.

Lemma 3. *The number of read placements of* expr_i *reads of read length* readLen*, such that there is exactly one contig of length* $\mathrm{tranLen}_i$*, is* $\binom{\mathrm{expr}_i-1}{\mathrm{tranLen}_i-\mathrm{readLen}}_{\mathrm{readLen}+1}$.

Proof. Recall that our definition of a read placement involves a fixed enumeration of the reads, from left to right. When we enumerate the reads in this way, each read can be thought to "newly cover" a certain number of bases, in the following sense: (i) read 1 newly covers all readLen of its bases, and (ii) for $r \geq 2$, read r newly covers all bases not covered by read $r-1$. For an example, see Figure 6. Let c_r be the number of bases newly covered by read r, and let $c = (c_1, c_2, \ldots, c_{\mathrm{expr}_i})$ be the "new-coverage pattern" for this read placement.

In the current lemma, we are only considering read placements that fully cover the transcript. There is a 1–1 correspondence between such read placements and new-coverage patterns. The reason for this is that the only way two read placements (based on the same fixed enumeration) can have the same new-coverage pattern is if there are gaps in the coverage of bases, and the lemma prohibits this. Thus the number of read placements which fully cover the transcript equals the number of distinct new-coverage patterns.

How many such new-coverage patterns are there? The first read always newly covers readLen bases, so we can ignore it. Each new-coverage pattern tells us how many of the remaining $\mathrm{tranLen}_i - \mathrm{readLen}$ bases are newly covered by the remaning $\mathrm{expr}_i - 1$ reads. Between 0 and readLen bases, inclusive, can be newly covered by each read. The bases are unlabelled, in the sense that the new-coverage pattern only (explicitly) says *how many* bases are covered by each read, not which particular bases are covered. (We can indirectly figure out which particular bases are covered, using the enumeration of reads, but since we fixed the enumeration in advance, there is only one possibility.) The reads are labelled, in the sense that the new-coverage pattern tells us how many bases each read newly contributes.

In other words, each new-coverage pattern tells us how to allocate $\mathrm{tranLen}_i - \mathrm{readLen}$ (unlabelled) balls among $\mathrm{expr}_i - 1$ (labelled) urns, allowing at most readLen balls to fall in each urn. As discussed in 3.3.3, the number of ways of doing this is $\binom{\mathrm{expr}_i-1}{\mathrm{tranLen}_i-\mathrm{readLen}}_{\mathrm{readLen}+1}$. ■

3.3.5 Contig placements

In addition to read placements, it is helpful also to think about "contig placements". A contig placement of $\mathrm{numContigsFromTran}_i$ contigs, each of length $\mathrm{contigLen}_j$ ($j \in \mathrm{contigsFromTran}_i$), within a transcript of length $\mathrm{tranLen}_i$, is defined to be a choice of start positions s_j, one for each contig $j \in \mathrm{contigsFromTran}_i$. The start positions are constrained by the definition of a contig: in any valid contig placement, there needs to be at least one uncovered base between each contig.

A basic question about contig placements is how many distinct contig placements exist. We will answer this question in two steps. First, suppose that we fix a particular ordering of the contigs, say, an ordering in which contig j starts (and hence ends) before contig j' whenever $j < j'$. How many contig placements satisfy this constraint?

Lemma 4. *The number of contig placements such that contig* j *starts before contig* j' *if and only if* $j < j'$ *is* $\binom{\mathrm{tranLen}_i+1-\sum_{j\in\mathrm{contigsFromTran}_i}\mathrm{contigLen}_j}{\mathrm{numContigsFromTran}_i}$.

Proof. Each contig has a (possibly) distinct length $\mathrm{contigLen}_j$. However, since contigs need to be disjoint (and, moreover, separated by an uncovered base), the contig lengths are only important insofar as they collectively reduce the effective length of the transcript. In other words, the number of contig placements is unchanged if (i) contig j is shortened by a certain number of bases δ_j, and (ii) the transcript is also shortened by δ_j bases. By using this fact repeatedly, the number of contig placements is unchanged if (i) for all j, contig j is shortened by $\mathrm{contigLen}_j$ bases, so that it now has length 0, and (ii) the transcript is shortened by $\sum_{j \in \mathrm{contigsFromTran}_i} \mathrm{contigLen}_j$.

we can simplify the situation by ignoring

imagining, equivalently to before, that (i) each contig has length 1, and (ii) each contig's original length $contigLen_j$ is subtracted from the transcript's length. Due to (i), we only need to choose where to place each empty space. The empty spaces are unlabelled, since

(so that we only count the empty spaces in between them)

In general, the number of ways to choose k unordered elements, with replacement, from a collection of size n, is $\binom{n+k-1}{k-1} = \binom{n+k-1}{n}$. The lemma follows from our previous remarks.

■

Lemma 5. *The number of ways to order* $\mathrm{numContigs}_i$ *contigs*

The number of "Only include placements that have the shortest contig first, then the next longest contig, etc. I.e., exclude different permutations of the contig relative orderings." is $\binom{n}{\mathbf{n}}$.

Proof. ■

3.3.6 Proof of the theorem

Lemma 6. *Under Stipulations 3 and 4 given above, the number of read placements such that there are (i) n contigs and (ii) $n_{c,l}$ contigs with length l and coverage c is:*

$$\mathrm{numPlacements}(\mathbf{n}; \mathrm{readLen}, \mathrm{tranLen}) = \binom{n}{\mathbf{n}} \binom{\mathrm{tranLen}_i + 1 - \sum_{j \in \mathscr{J}} \mathrm{contigLen}_j}{n} \prod_{j \in \mathscr{J}} \binom{\mathrm{coverage}_j - 1}{\mathrm{contigLen}_j - \mathrm{readLen}}_{\mathrm{readLen}+1}$$

Lemma 7. *Under Stipulations 3 and 4 given above, the total number of possible read placements is*

$$\mathrm{numPlacements}(\mathrm{readLen}, \mathrm{tranLen}) = \binom{\mathrm{expr}_i + \mathrm{tranLen}_i - \mathrm{readLen}}{\mathrm{expr}_i}$$

If, for example, there are two balls in urn 1, the identity of the two balls doesn't matter - the situation is identical if "the first" two balls fell in urn 1 or if some other two balls fell in urn 1. On the other hand, the situation is distinct if there are two balls in urn 1 versus two balls in urn 2.

of distinct ways in which $k := \mathrm{contigLen}'_j - \mathrm{readLen}$ balls (i.e., bases) can be allocated to $n := \mathrm{coverage}_j - 1$ urns (i.e., reads), allowing at most $N - 1 := \mathrm{readLen}$ balls (i.e., bases) to fall in each urn (i.e., read). It is known that the number of ways to do this is $\binom{n}{k}_N$: see comments by N-E. Fahssi from OEIS entries `http://oeis.org/A008287` and `http://oeis.org/A035343`, and cf. also Fahssi (2012).

Claim: $\text{numArrangements}(\text{coverage}_j, \text{contigLen}'_j, \text{readLen}) = \binom{\text{coverage}_j - 1}{\text{contigLen}'_j - \text{readLen}}_{\text{readLen}+1}$, where

$$\binom{n}{k}_N = \begin{cases} 0 & \text{if } n < 0 \text{ or } k < 0 \text{ or } k > (N-1) \cdot n \\ 1 & \text{if } n = 0 \text{ and } k = 0 \\ \sum_{j=0}^{N-1} \binom{n-1}{k-j}_N & \text{otherwise} \end{cases}$$

Proof: Fix an enumeration of the coverage_j reads involved in contig j from left to right, as described in item (b) above. Each read can be thought to "newly contribute" a certain number of bases, in the sense that no previous (in the enumeration) read covers these bases, but the read in question does cover them. See Figure 6 for an example. The first (and leftmost) read always contributes readLen bases. Each remaining read can contribute between 0 bases (if it completely overlaps the previous read) and readLen bases (if it starts just after the previous read), assuming that we require window size of 0.

For each arrangement a of the reads, let $x = x(a)$ be the corresponding "new-contribution" vector, i.e., x_r is the number of bases newly contributed by read r in the arrangement. Note that there is a 1-1 correspondence between arrangements a and new-contribution vectors x. Thus the number of arrangements of the reads into a contig of length $\text{contigLen}'_j$ is equal to the number of valid new-contribution vectors, i.e., the number of vectors $x \in \{\text{readLen}\} \times \{0, \ldots, \text{readLen}\}^{\text{coverage}_j - 1}$ such that $\sum_{r=1}^{\text{coverage}_j - 1} x_r = \text{contigLen}'_j$.

How many such new-contribution vectors are there? The first read always newly covers readLen bases. Each new-contribution vector says how to allocate the remaining $\text{contigLen}'_j - \text{readLen}$ bases among the remaining $\text{coverage}_j - 1$ reads, allowing at most readLen positions to be allocated to each read. We can think of the bases as being unlabelled, in the sense that the new-contribution vectors do not keep track of *which* specific (identified) bases are newly covered by each read, but rather just *how many* bases are newly covered by each read. We can think of the reads as being labelled, in the sense that each new-contribution vector says how many bases are newly contributed by each read which is labelled by its position within the (fixed) enumeration. Thus, the number of new-contribution vectors is the number of distinct ways in which $k := \text{contigLen}'_j - \text{readLen}$ balls (i.e., bases) can be allocated to $n := \text{coverage}_j - 1$ urns (i.e., reads), allowing at most $N - 1 := \text{readLen}$ balls (i.e., bases) to fall in each urn (i.e., read). It is known that the number of ways to do this is $\binom{n}{k}_N$: see comments by N-E. Fahssi from OEIS entries `http://oeis.org/A008287` and `http://oeis.org/A035343`, and cf. also Fahssi (2012).

For empirical evidence of the claim, see `howManyWaysToCoverAContig.py`. ■

Claim: $\sum_{l=\text{readLen}}^{\text{readLen}\cdot\text{coverage}_j} \text{numArrangements}(\text{coverage}_j, l, \text{readLen}) = (\text{readLen}+1)^{\text{coverage}_j - 1}$.

Proof: Note that

$$\begin{aligned} &\sum_{l=\text{readLen}}^{\text{readLen}\cdot\text{coverage}_j} \text{numArrangements}(\text{coverage}_j, l, \text{readLen}) \\ = &\sum_{l=\text{readLen}}^{\text{readLen}\cdot\text{coverage}_j} \binom{\text{coverage}_j - 1}{l - \text{readLen}}_{\text{readLen}+1} \\ = &\sum_{l=0}^{\text{readLen}\cdot(\text{coverage}_j - 1)} \binom{\text{coverage}_j - 1}{l}_{\text{readLen}+1} \end{aligned}$$

the $(\text{coverage}_j - 1)$th row sum of the array with entries $(n,k) = \binom{n}{k}_{\text{readLen}+1}$. It is a fact (see `http://oeis.org/A027907` for the trinomial case) that these row sums are $n \mapsto N^n$, i.e., in our case, the $(\text{coverage}_j - 1)$th row sum is $(\text{readLen} + 1)^{\text{coverage}_j - 1}$.

For empirical evidence, see `howManyWaysToCoverAContig.py`. ■

3.3.7 The distribution

Let ContigCounts_i be as defined in Assumption 2. Note that $\sum_l \sum_c \text{ContigCounts}_{i,l,c} = \text{NumContigsFromTran}_i$.

Let $\text{numPlacements}(\text{contigCounts}_i; \text{tranLen}_i, \text{readLen})$ be the number of ways to place $\text{numContigsFromTran}_i$ reads

3.3.8 Proof

specific assignments

A key thing is how to represent the contigs? The answer is that:

1. There are transcripts Tran_i, $i \in \{1, 2, \ldots, \text{NumTrans}\}$.
2. For each transcript i, there are contigs $\text{Contig}_{i,j}$, $j \in \{1, 2, \ldots, \text{NumContigsFromTran}_i\}$.

Thus:

$$\begin{aligned}
&P(\text{contigs}, \text{coverage} | \text{trans}, \text{expr}) \\
&= \prod_{i=1}^{\text{numTrans}} P(\text{contigs}_i, \text{coverage}_i | \text{tran}_i, \text{expr}_i)
\end{aligned}$$

Note that we do not know the assignment of contigs to transcripts. So we also use the representation:

1. There are transcripts Tran_i, $i \in \{1, 2, \ldots, \text{NumTrans}\}$.
2. There are contigs Contig_j, $j \in \{1, 2, \ldots, \text{NumContigs}\}$.
3. There is a transcript-to-contigs map $\text{ContigsFromTran}_i \subset \{1, 2, \ldots, \text{NumContigs}\}$, with $\text{ContigsFromTran}_i \cap \text{ContigsFromTran}_{i'} = \varnothing$ and $\cup_i \text{ContigsFromTran}_i = \{1, 2, \ldots, \text{NumContigs}\}$.

In this representation, one needs to account for the different ways to assign contig indices to transcripts. However, we consider all of these to be equivalent so there is no need to do further normalization. Thus

$$\begin{aligned}
&P(\text{contigs}, \text{coverage}, \text{contigsFromTrans} | \text{trans}, \text{expr}) \\
&= P(\text{contigs}, \text{coverage} | \text{trans}, \text{expr}) \\
&= \prod_{i=1}^{\text{numTrans}} P(\text{contigs}_i, \text{coverage}_i | \text{tran}_i, \text{expr}_i)
\end{aligned}$$

3.3.9 The distribution of $\text{Contigs}_i, \text{Coverage}_i | \text{Tran}_i, \text{Expr}_i$

In this section, all statements are understood to be asserted to hold, conditionally on $\{\omega : \text{Tran}_i = \text{tran}_i, \text{Expr}_i = \text{expr}_i\}$.

We use the following shorthand notation:

1. $n = \text{numContigsFromTran}_i$
2. $\mathscr{J} = \text{contigsFromTran}_i$.

```
transscript:   01234567890123456789
arrangement 1: aaaaaaaaabbbbbb
arrangement 2: aaaaaaaaa bbbbbb
arrangement 3:    bbbbbb aaaaaaaaa
etc
```

Figure 1: Illustration of our assumption that J is constant with respect to contig ordering and start positions within a transcript. In this case, regardless of how we arrange the two contigs `aaaaaaaaa` and `bbbbbb` within the transcript, we assume that J is constant.

3. $n_{c,l} = |\{j \in \mathscr{J} : \text{coverage}_j = c, \text{contigLen}_j = l\}|$.
4. $\mathbf{n} = (n_{c,l})$, thought of as a vector or matrix as appropriate.

Note that $\sum_c \sum_l n_{c,l} = n = |\mathscr{J}|$.

What is $P(\text{contigs}_i, \text{coverage}_i | \text{tran}_i, \text{expr}_i)$? We find the exact distribution under the following stipulations:

1. Stipulation 1: The start positions of the expr_i reads are distributed uniformly among the possible start positions.
2. Stipulation 2: The start positions of the expr_i reads are independent.
3. Stipulation 3: All reads share a common read length readLen.
4. Stipulation 4: A contig is defined to be a maximal contiguous subsequence of the transcript such that every base is covered by at least one read.

Lemma 8. *Under Stipulations 3 and 4 given above, the number of read placements such that there are (i) n contigs and (ii) $n_{c,l}$ contigs with length l and coverage c is:*

$$\text{numPlacements}(\mathbf{n}; \text{readLen}, \text{tranLen}) = \binom{n}{\mathbf{n}} \binom{\text{tranLen}_i + 1 - \sum_{j \in \mathscr{J}} \text{contigLen}_j}{n} \prod_{j \in \mathscr{J}} \binom{\text{coverage}_j - 1}{\text{contigLen}_j - \text{readLen}}_{\text{readLen}+1}$$

Lemma 9. *Under Stipulations 3 and 4 given above, the total number of possible read placements is*

$$\text{numPlacements}(\text{readLen}, \text{tranLen}) = \binom{\text{expr}_i + \text{tranLen}_i - \text{readLen}}{\text{expr}_i}$$

Theorem 10. *Under the stipulations given above, the*

$$P(\text{contigs}_i, \text{coverage}_i | \text{tran}_i, \text{expr}_i) = \frac{\text{numPlacements}(\mathbf{n}; \text{readLen}, \text{tranLen})}{\text{numPlacements}(\text{readLen}, \text{tranLen})}$$

$$= \frac{\binom{n}{\mathbf{n}} \binom{\text{tranLen}_i + 1 - \sum_{j \in \mathscr{J}} \text{contigLen}_j}{n} \prod_{j \in \mathscr{J}} \binom{\text{coverage}_j - 1}{\text{contigLen}_j - \text{readLen}}_{\text{readLen}+1}}{\binom{\text{expr}_i + \text{tranLen}_i - \text{readLen}}{\text{expr}_i}}$$

4 Mode of the joint distribution

We want to maximize (or at least approximately maximize) the function

$$J(\text{trans}, \text{expr}, \text{coverage}, \text{contigsFromTrans}) = P(\text{reads}, \text{trans}, \text{expr}, \text{contigs}, \text{coverage}, \text{contigsFromTrans})$$

with $(\text{reads}, \text{contigs})$ held fixed.

We will assume that the contigs are "correct", i.e., are error-free contiguous subsequences of the transcripts. We will also assume that J is constant with respect to contig ordering and start positions within a transcript; this is exactly true for everything except for the data likelihood, and for the RSEM-based likelihood it is almost true (or is it exactly true?); for an example, see Figure 1.

Due to these assumptions, we do not need to explicitly write down the sequences of the transcripts that maximize the objective. Rather, we just need to answer the questions: (i) Which contigs should be said to come from the same transcripts as each other, and (ii) how long should these transcripts be? so as to maximize the objective J.

We also need to find the optimal coverage of the contigs. This, together with the assignment of contigs to transcripts, will automatically give us the optimal expression of the transcripts.

We will use the following approach based on greedy coordinate ascent.

Step 1: Run RSEM, using the contigs as the reference. This will give us τ_j, the MLE of the fraction of reads that come from each contig.

Find the optimal coverage coverage*.

1. Start with an RSEM-based estimate coverage* of the coverage, based on the contigs and the reads. (coverage$^* = \lfloor \tau \cdot \text{numReads} \rfloor$ where τ is the "transcript-level expression" (based on contigs) produced by RSEM.)
2. Figure out analytically which contigs should be joined into transcripts, based on the coverage and length of the contigs. This gives trans* and contigsFromTrans*. (We never need to know explicitly the sequence of the transcripts outside of the contigs.)
3. Sum appropriate entries of coverage* to get expr*.

The tricky part is item 2. In the following two subsections, we will (i) determine how the data likelihood $P(\text{reads}|\text{trans},\text{expr})$ changes when we join contigs into a transcript, and (ii) determine how the overall joint probability changes when we join contigs into a transcript. After that, we will (iii) formulate a simple algorithm to make high-probability transcripts from contigs.

should be "joined together" into transcripts, and (ii) how long should the transcripts be? so as to maximize the objective J.

Due to these assumptions, maximization of J requires us to answer the following question: Which contigs should be "joined together" into transcripts so as to maximize the objective J?

Due to these assumptions, maximization of J involves answering two (linked) questions:

1. Which contigs should be "joined together" into transcripts so as to maximize the objective J?

   Order doesn't matter here. A transcript could be longer than the corresponding contigs.
2. What are the expression and coverage? A transcript's expression is the sum of the coverage of its contigs.

We want an approximation of the mode that can be found as easily as possible. Our strategy is:

1. Start with an RSEM-based estimate coverage* of the coverage, based on the contigs and the reads. (coverage$^* = \lfloor \tau \cdot \text{numReads} \rfloor$ where τ is the "transcript-level expression" (based on contigs) produced by RSEM.)
2. Figure out analytically which contigs should be joined into transcripts, based on the coverage and length of the contigs. This gives trans* and contigsFromTrans*. (We never need to know explicitly the sequence of the transcripts outside of the contigs.)
3. Sum appropriate entries of coverage* to get expr*.

The tricky part is item 2. In the following two subsections, we will (i) determine how the data likelihood $P(\text{reads}|\text{trans},\text{expr})$ changes when we join contigs into a transcript, and (ii) determine how the overall joint probability changes when we join contigs into a transcript. After that, we will (iii) formulate a simple algorithm to make high-probability transcripts from contigs.

```
trans:  1:  AAAAAAAAAAA
        2:  GGGGGGG
        ... [other transcripts]
trans': 12: TTTAAAAAAAAAAATTTTTTTTGGGGGGGTTTT
        ... [other transcripts]
```

Figure 2: Here we join transcripts 1 and 2 into transcript 12, with extra bases (all "T"s in this case) in between and on the ends.

4.1 How does the joint distribution change when we join two transcripts together?

Throughout this subsection, we consider two transcript sets:

1. trans, which contains transcripts with indices $\mathscr{I} \cup \{1,2\}$, and
2. trans$'$, which contains transcripts with indices $\mathscr{I} \cup \{12\}$.

Here, transcript 12 contains transcript 1 and transcript 2 as disjoint subsequences, with possibly some extra bases added in between 1's and 2's sequences or on the ends, but not within 1's or 2's sequences. See Figure 2. The set $\mathscr{I}$ contains the indices of transcripts common to both trans and trans$'$. Obviously, $\text{numTrans}' = \text{numTrans} - 1$. Also, $\text{expr}'_{12} = \text{expr}_1 + \text{expr}_2$. Also, $\text{numContigsFromTran}'_{12} = \text{numContigsFromTran}_1 + \text{numContigsFromTran}_2$ and $\text{contigsFromTran}'_{12} = \text{contigsFromTran}_1 \cup \text{contigsFromTran}_2$.

The idea, of course, is that trans is the "original transcripts" and trans$'$ is the "hopefully higher-probability transcripts". Initially, trans will be the contigs.

4.1.1 How does the data likelihood $P(\text{reads}|\text{trans},\text{expr})$ change when we join two transcripts together?

Recall that $P(\text{reads}|\text{trans},\text{expr}) = P_{RSEM}(\text{reads}|\text{trans},\text{expr})$. We use the notation $\mathbb{P}(\dots) = P_{RSEM}(\dots|\text{trans},\text{expr})$ and $\mathbb{P}'(\dots) = P_{RSEM}(\dots|\text{trans}',\text{expr}')$. We want to figure out how $\mathbb{P}$ differs from $\mathbb{P}'$.

According to RSEM's model, the joint distribution decomposes as

$$\prod_{n=1}^{N} \mathbb{P}(g_n, f_n, s_n, o_n, l_n, q_n, r_n) = \prod_{n=1}^{N} \mathbb{P}(g_n)\mathbb{P}(f_n|g_n)\mathbb{P}(s_n|g_n, f_n)\mathbb{P}(l_n|f_n)\mathbb{P}(o_n|g_n)\mathbb{P}(q_n)\mathbb{P}(r_n|g_n, f_n, s_n, l_n, o_n, q_n)$$

where (assuming single-end data)

1. $N = \text{numReads}$, indexed by n,
2. G_n is the parent transcript's index,
3. F_n is the fragment length,
4. S_n is the start position,
5. O_n is the orientation,
6. L_n is the read length,
7. Q_n is the quality score sequence, and
8. R_n is the read sequence.

When the parent transcript $G_n = g_n \in \mathscr{I}$, the shared transcript index set, then $\mathbb{P}(g_n, f_n, s_n, o_n, l_n, q_n, r_n) = \mathbb{P}'(g_n, f_n, s_n, o_n, l_n, q_n, r_n)$, since $\mathbb{P}(g_n) = \mathbb{P}'(g_n)$ and all the other parts of the decomposition are also the same.

What about when G_n is 1 or 2? In that case, $G'_n = 12$. So:

1. G_n: $\mathbb{P}'(G_n = 12) = \mathbb{P}(G_n = 1) + \mathbb{P}(G_n = 2)$, since every read that came from 1 or 2 under $\mathbb{P}$ comes from 12 under $\mathbb{P}'$, and no other reads come from $\mathbb{P}'$. Specifically, $\mathbb{P}(G_n = j) = \text{expr}_j/\text{numReads}$, and $\mathbb{P}'(G_n = j) = \text{expr}'_j/\text{numReads}$.
2. F_n: It is difficult to say how the fragment length distribution $F_n|G_n$ changes, since it is modelled nonparametrically. However, it seems perhaps reasonable to assume, in our present analysis, that $\mathbb{P}(f_n|g_n) = 1(f_n = \text{tranLen}_{g_n})$, i.e., to pretend that every fragment consists of the entire transcript (and hence to ignore the fact that reads come from fragments, not transcripts). This is similar to the first RSEM paper.
3. S_n: For the start position S_n, we will assume a uniform RSPD along the whole feasible length of the fragment, i.e., the transcript. So:

$$\mathbb{P}(s_n|g_n, f_n) = (\text{tranLen}_{g_n} - \text{readLen} + 1)^{-1}$$
$$\mathbb{P}'(s_n|g_n, f_n) = (\text{tranLen}'_{g_n} - \text{readLen} + 1)^{-1}$$

4. O_n: The orientation's conditional distribution is the same under $\mathbb{P}$ and $\mathbb{P}'$.
5. L_n: The read length's conditional distribution is the same under $\mathbb{P}$ and $\mathbb{P}'$. We will assume that the read length is fixed at readLen.
6. Q_n: The quality score's conditional distribution is the same under $\mathbb{P}$ and $\mathbb{P}'$.
7. R_n: The read sequence's conditional distribution is the same under $\mathbb{P}$ and $\mathbb{P}'$.

Putting this all together, the odds ratio of the RSEM joint distributions is

$$\begin{aligned}
\frac{\prod_{n=1}^N \mathbb{P}(g_n, f_n, s_n, o_n, l_n, q_n, r_n)}{\prod_{n=1}^N \mathbb{P}'(g'_n, f_n, s_n, o_n, l_n, q_n, r_n)} &= \frac{\prod_{n:g_n\in\{1,2\}} \mathbb{P}(g_n)(\text{tranLen}_{g_n} - \text{readLen} + 1)^{-1}}{\prod_{n:g'_n=12} \mathbb{P}'(g'_n)(\text{tranLen}'_{g'_n} - \text{readLen} + 1)^{-1}} \\
&= \frac{\prod_{j\in\{1,2\}} \left[(\text{expr}_j/\text{numReads})(\text{tranLen}_j - \text{readLen} + 1)^{-1}\right]^{|\{n:g_n=j\}|}}{\left[(\text{expr}'_{12}/\text{numReads})(\text{tranLen}'_{12} - \text{readLen} + 1)^{-1}\right]^{|\{n:g'_n=12\}|}} \\
&= \frac{\prod_{j\in\{1,2\}} \left[\text{expr}_j \cdot (\text{tranLen}_j - \text{readLen} + 1)^{-1}\right]^{|\{n:g_n=j\}|}}{\left[\text{expr}'_{12} \cdot (\text{tranLen}'_{12} - \text{readLen} + 1)^{-1}\right]^{|\{n:g'_n=12\}|}} \\
&= \frac{\prod_{j\in\{1,2\}} \left[\text{expr}_j \cdot (\text{tranLen}_j - \text{readLen} + 1)^{-1}\right]^{\text{expr}_j}}{\left[\text{expr}'_{12} \cdot (\text{tranLen}'_{12} - \text{readLen} + 1)^{-1}\right]^{\text{expr}'_{12}}}
\end{aligned}$$

where $g'_n = 12$ if $g_n \in \{1,2\}$, and $g'_n = g_n$ otherwise. The penultimate identity holds because numReads does not depend on g_n. The last identity holds because $|\{n : g_n = j\}| = \text{expr}_j$ and $|\{n : g'_n = 12\}| = \text{expr}_{12}$.

Of course, we are really interested in the log ratio of the marginal (wrt RSEM) distributions $P(\text{reads}|\text{trans}, \text{expr})$ and $P(\text{reads}|\text{trans}', \text{expr}')$. We will approximate the integrals via BIC:

$$\begin{aligned}
\log P(\text{reads}|\text{trans}, \text{expr}) &= \log \mathbb{P}(l, r, q) \\
&= \log \int \prod_{n=1}^N \mathbb{P}(g_n, f_n, s_n, o_n, l_n, q_n, r_n)\, dg\, df\, ds\, do \\
&= \Big(\sum_{n=1}^N \log \mathbb{P}(g^*_n, f^*_n, s^*_n, o^*_n, l_n, q_n, r_n) \Big) - \frac{1}{2} \text{numTrans} \log \text{numReads}
\end{aligned}$$

So the log odds is:

$$\begin{aligned}
&\log P(\text{reads}|\text{trans},\text{expr}) - \log P(\text{reads}|\text{trans}',\text{expr}') \\
&= \log \mathbb{P}(l,r,q) - \log \mathbb{P}'(l,r,q) \\
&= \Big(\sum_{n=1}^{N} \log \mathbb{P}(g_n^*, f_n^*, s_n^*, o_n^*, l_n, q_n, r_n)\Big) - \frac{1}{2}\text{numTrans} \log \text{numReads} \\
&\qquad - \Big(\sum_{n=1}^{N} \log \mathbb{P}(g_n'^*, f_n^*, s_n^*, o_n^*, l_n, q_n, r_n)\Big) + \frac{1}{2}\text{numTrans}' \log \text{numReads} \\
&= \log \Big(\frac{\prod_{n=1}^{N} \mathbb{P}(g_n^*, f_n^*, s_n^*, o_n^*, l_n, q_n, r_n)}{\prod_{n=1}^{N} \mathbb{P}(g_n'^*, f_n^*, s_n^*, o_n^*, l_n, q_n, r_n}\Big) - \frac{1}{2} \log \text{numReads}
\end{aligned}$$

We have used that $\text{numTrans} - \text{numTrans}' = 1$. Substituting from the joint log odds above,

$$\begin{aligned}
&\log P(\text{reads}|\text{trans},\text{expr}) - \log P(\text{reads}|\text{trans}',\text{expr}') \\
&= \log \left(\frac{\prod_{j\in\{1,2\}} \left[\text{expr}_j \cdot (\text{tranLen}_j - \text{readLen} + 1)^{-1}\right]^{\text{expr}_j}}{\left[\text{expr}'_{12} \cdot (\text{tranLen}'_{12} - \text{readLen} + 1)^{-1}\right]^{\text{expr}'_{12}}} \right) - \frac{1}{2} \log \text{numReads} \\
&= -\left(\text{expr}'_{12} \log \text{expr}'_{12} - \sum_{j\in\{1,2\}} \text{expr}_j \log \text{expr}_j \right) \\
&\quad + \left(\text{expr}'_{12} \log(\text{tranLen}'_{12} - \text{readLen} + 1) - \sum_{j\in\{1,2\}} \text{expr}_j \log(\text{tranLen}_j - \text{readLen} + 1) \right) \\
&\quad - \frac{1}{2} \log \text{numReads}
\end{aligned}$$

4.1.2 How does the prior $P(\text{trans},\text{expr})$ change when we join two transcripts together?

Recall that

$$P(\text{trans},\text{expr}) = \frac{(\mu_0(1-p_0)^{r_0}/\text{numReads})^{\text{numTrans}}}{\text{numTrans}!} e^{-\mu_0} \prod_{i=1}^{\text{numTrans}} \binom{\text{tranLen}_i + r_0 - 1}{\text{tranLen}_i} (p_0/4)^{\text{tranLen}_i}$$

so the log odds

$$\begin{aligned}
&\log P(\text{trans},\text{expr}) - \log P(\text{trans}',\text{expr}') \\
&= \log \frac{\frac{(\mu_0(1-p_0)^{r_0}/\text{numReads})^{\text{numTrans}}}{\text{numTrans}!} e^{-\mu_0} \prod_{i=1}^{\text{numTrans}} \binom{\text{tranLen}_i + r_0 - 1}{\text{tranLen}_i} (p_0/4)^{\text{tranLen}_i}}{\frac{(\mu_0(1-p_0)^{r_0}/\text{numReads})^{\text{numTrans}'}}{\text{numTrans}'!} e^{-\mu_0} \prod_{i=1}^{\text{numTrans}'} \binom{\text{tranLen}'_i + r_0 - 1}{\text{tranLen}'_i} (p_0/4)^{\text{tranLen}'_i}}
\end{aligned}$$

Note that $\text{tranLen}_i = \text{tranLen}'_i$ for all $i \notin \{1,2,12\}$, so there is a lot of cancellation:

$$\cdots = \log \left(\frac{\frac{(\mu_0(1-p_0)^{r_0}/\text{numReads})^{\text{numTrans}}}{\text{numTrans}!}}{\frac{(\mu_0(1-p_0)^{r_0}/\text{numReads})^{\text{numTrans}'}}{\text{numTrans}'!}} \cdot \frac{\prod_{i\in\{1,2\}} \binom{\text{tranLen}_i + r_0 - 1}{\text{tranLen}_i} (p_0/4)^{\text{tranLen}_i}}{\binom{\text{tranLen}'_{12} + r_0 - 1}{\text{tranLen}'_{12}} (p_0/4)^{\text{tranLen}'_{12}}} \right)$$

Note that $\text{numTrans} - \text{numTrans}' = 1$ and $\text{numTrans}!/\text{numTrans}'! = \text{numTrans}$. So

$$\cdots = \log\left(\frac{(\mu_0(1-p_0)^{r_0}/\text{numReads})^{\text{numTrans}-\text{numTrans}'}}{\text{numTrans}!/\text{numTrans}'!} \cdot \frac{\prod_{i\in\{1,2\}} \binom{\text{tranLen}_i + r_0 - 1}{\text{tranLen}_i} (p_0/4)^{\text{tranLen}_i}}{\binom{\text{tranLen}'_{12} + r_0 - 1}{\text{tranLen}'_{12}} (p_0/4)^{\text{tranLen}'_{12}}}\right)$$

$$= \log\left(\frac{\mu_0(1-p_0)^{r_0}}{\text{numReads}\cdot\text{numTrans}} \cdot \frac{\prod_{i\in\{1,2\}} \binom{\text{tranLen}_i + r_0 - 1}{\text{tranLen}_i} (p_0/4)^{\text{tranLen}_i}}{\binom{\text{tranLen}'_{12} + r_0 - 1}{\text{tranLen}'_{12}} (p_0/4)^{\text{tranLen}'_{12}}}\right)$$

Applying the log,

$$\begin{aligned}\log P(\text{trans},\text{expr}) - \log P(\text{trans}',\text{expr}') = {} & \log\left(\frac{\mu_0(1-p_0)^{r_0}}{\text{numReads}\cdot\text{numTrans}}\right) \\ & + \left(-\log\binom{\text{tranLen}'_{12} + r_0 - 1}{\text{tranLen}'_{12}} + \sum_{i\in\{1,2\}} \log\binom{\text{tranLen}_i + r_0 - 1}{\text{tranLen}_i}\right) \\ & + \left(-\text{tranLen}'_{12} + \sum_{i\in\{1,2\}} \text{tranLen}_i\right)\log(p_0/4)\end{aligned}$$

4.1.3 How does $P(\text{contig},\text{coverage},\text{contigsFromTrans}|\text{trans},\text{expr})$ change when we join two transcripts together? Try 2

Note that

$$\begin{aligned}&\frac{P(\text{contig},\text{coverage},\text{contigsFromTrans}'|\text{trans}',\text{expr}')}{P(\text{contig},\text{coverage},\text{contigsFromTrans}|\text{trans},\text{expr})} \\ &= \frac{P(\text{contig}'_{(12)},\text{coverage}'_{(12)}|\text{trans}'_{12},\text{expr}'_{12})}{P(\text{contig}_{(1)},\text{coverage}_{(1)}|\text{trans}_1,\text{expr}_1)P(\text{contig}_{(2)},\text{coverage}_{(2)}|\text{trans}_2,\text{expr}_2)} \\ &= \frac{\dfrac{\binom{\text{numContigsFromTran}'_{12}}{\text{numContigsWithStats}'_{12}}\binom{\text{tranLen}_i+1-\sum_{j\in\text{contigsFromTran}'_{12}}\text{contigLen}_j}{\text{numContigsFromTran}'_{12}}\prod_{j\in\text{contigsFromTran}'_{12}}{\binom{\text{coverage}_j-1}{\text{contigLen}_j-\text{readLen}}}_{\text{readLen}+1}}{\binom{\text{expr}'_{12}+\text{tranLen}'_{12}-\text{readLen}}{\text{expr}'_{12}}}}{\prod_{i\in\{1,2\}}\dfrac{\binom{\text{numContigsFromTran}_i}{\text{numContigsWithStats}_i}\binom{\text{tranLen}_i+1-\sum_{j\in\text{contigsFromTran}_i}\text{contigLen}_j}{\text{numContigsFromTran}_i}\prod_{j\in\text{contigsFromTran}_i}{\binom{\text{coverage}_j-1}{\text{contigLen}_j-\text{readLen}}}_{\text{readLen}+1}}{\binom{\text{expr}_i+\text{tranLen}_i-\text{readLen}}{\text{expr}_i}}}\end{aligned}$$

where $\text{numContigsWithStats}_i = (\text{numContigsWithStats}_{i,c,l})$, $\text{numContigsWithStats}_{i,c,l} = |\{j \in \text{contigsFromTran}_i : \text{coverage}_j = c, \text{contigLen}_j = l\}|$. Note that $\text{contigsFromTran}'_{12} = \text{contigsFromTran}_1 \dot\cup \text{contigsFromTran}_2$ so the rightmost products of factors cancel and we get:

$$\begin{aligned}&= \frac{\dfrac{\binom{\text{numContigsFromTran}'_{12}}{\text{numContigsWithStats}'_{12}}\binom{\text{tranLen}'_{12}+1-\sum_{j\in\text{contigsFromTran}'_{12}}\text{contigLen}_j}{\text{numContigsFromTran}'_{12}}}{\binom{\text{expr}'_{12}+\text{tranLen}'_{12}-\text{readLen}}{\text{expr}'_{12}}}}{\prod_{i\in\{1,2\}}\dfrac{\binom{\text{numContigsFromTran}_i}{\text{numContigsWithStats}_i}\binom{\text{tranLen}_i+1-\sum_{j\in\text{contigsFromTran}_i}\text{contigLen}_j}{\text{numContigsFromTran}_i}}{\binom{\text{expr}_i+\text{tranLen}_i-\text{readLen}}{\text{expr}_i}}} \\ &= \rho_1\rho_2\rho_3\end{aligned}$$

We consider each of the three factors separately.

First,

$$\begin{aligned}
\rho_1 &= \frac{\binom{\text{numContigsFromTran}'_{12}}{\text{numContigsWithStats}'_{12}}}{\prod_{i\in\{1,2\}}\binom{\text{numContigsFromTran}_i}{\text{numContigsWithStats}_i}} \\
&= \frac{\dfrac{\text{numContigsFromTran}'_{12}!}{\prod_c\prod_l \text{numContigsWithStats}'_{12,c,l}!}}{\dfrac{\text{numContigsFromTran}_1!\text{numContigsFromTran}_2!}{\prod_c\prod_l \text{numContigsWithStats}_{1,c,l}!\text{numContigsWithStats}_{2,c,l}!}}
\end{aligned}$$

Note that $\text{numContigsFromTran}'_{12} = \text{numContigsFromTran}_1 + \text{numContigsFromTran}_2$ and $\text{numContigsWithStats}'_{12,c,l} = \text{numContigsWithStats}_{1,c,l} + \text{numContigsWithStats}_{2,c,l}$, so we can rearrange:

$$\begin{aligned}
\rho_1 &= \frac{\dfrac{\text{numContigsFromTran}'_{12}!}{\text{numContigsFromTran}_1!\text{numContigsFromTran}_2!}}{\prod_c\prod_l \dfrac{\text{numContigsWithStats}'_{12,c,l}!}{\text{numContigsWithStats}_{1,c,l}!\text{numContigsWithStats}_{2,c,l}!}} \\
&= \frac{\binom{\text{numContigsFromTran}'_{12}}{\text{numContigsFromTran}_1}}{\prod_c\prod_l\binom{\text{numContigsWithStats}'_{12,c,l}}{\text{numContigsWithStats}_{1,c,l}}}
\end{aligned}$$

Second,

$$\begin{aligned}
\rho_2 &= \frac{\binom{\text{tranLen}'_{12}+1-\sum_{j\in\text{contigsFromTran}'_{12}}\text{contigLen}_j}{\text{numContigsFromTran}'_{12}}}{\prod_{i\in\{1,2\}}\binom{\text{tranLen}_i+1-\sum_{j\in\text{contigsFromTran}_i}\text{contigLen}_j}{\text{numContigsFromTran}_i}} \\
&= \frac{\dfrac{(\text{tranLen}'_{12}+1-\sum_{j\in\text{contigsFromTran}'_{12}}\text{contigLen}_j)!}{\text{numContigsFromTran}'_{12}!(\text{tranLen}'_{12}+1-\sum_{j\in\text{contigsFromTran}'_{12}}\text{contigLen}_j-\text{numContigsFromTran}'_{12})!}}{\prod_{i\in\{1,2\}}\dfrac{(\text{tranLen}_i+1-\sum_{j\in\text{contigsFromTran}_i}\text{contigLen}_j)!}{\text{numContigsFromTran}_i!(\text{tranLen}_i+1-\sum_{j\in\text{contigsFromTran}_i}\text{contigLen}_j-\text{numContigsFromTran}_i)!}} \\
&= \frac{1}{\binom{\text{numContigsFromTran}'_{12}}{\text{numContigsFromTran}_1}}\frac{\dfrac{(\text{tranLen}'_{12}+1-\sum_{j\in\text{contigsFromTran}'_{12}}\text{contigLen}_j)!}{(\text{tranLen}'_{12}+1-\sum_{j\in\text{contigsFromTran}'_{12}}\text{contigLen}_j-\text{numContigsFromTran}'_{12})!}}{\prod_{i\in\{1,2\}}\dfrac{(\text{tranLen}_i+1-\sum_{j\in\text{contigsFromTran}_i}\text{contigLen}_j)!}{(\text{tranLen}_i+1-\sum_{j\in\text{contigsFromTran}_i}\text{contigLen}_j-\text{numContigsFromTran}_i)!}} \\
&= \frac{\dfrac{(\text{tranLen}'_{12}+1-\sum_{j\in\text{contigsFromTran}'_{12}}\text{contigLen}_j)_{\underline{(\text{numContigsFromTran}'_{12})}}}{\prod_{i\in\{1,2\}}(\text{tranLen}_i+1-\sum_{j\in\text{contigsFromTran}_i}\text{contigLen}_j)_{\underline{(\text{numContigsFromTran}_i)}}}}{\binom{\text{numContigsFromTran}'_{12}}{\text{numContigsFromTran}_1}}
\end{aligned}$$

where $(n)_{\underline{(k)}} = n!/(n-k)!$.

Third,

$$
\begin{aligned}
\rho_3 &= \frac{\frac{1}{\binom{\mathrm{expr}'_{12}+\mathrm{tranLen}'_{12}-\mathrm{readLen}}{\mathrm{expr}'_{12}}}}{\prod_{i\in\{1,2\}} \frac{1}{\binom{\mathrm{expr}_i+\mathrm{tranLen}_i-\mathrm{readLen}}{\mathrm{expr}_i}}} \\
&= \frac{\prod_{i\in\{1,2\}} \binom{\mathrm{expr}_i+\mathrm{tranLen}_i-\mathrm{readLen}}{\mathrm{expr}_i}}{\binom{\mathrm{expr}'_{12}+\mathrm{tranLen}'_{12}-\mathrm{readLen}}{\mathrm{expr}'_{12}}} \\
&= \frac{\prod_{i\in\{1,2\}} \frac{(\mathrm{expr}_i+\mathrm{tranLen}_i-\mathrm{readLen})!}{\mathrm{expr}_i!(\mathrm{tranLen}_i-\mathrm{readLen})!}}{\frac{(\mathrm{expr}'_{12}+\mathrm{tranLen}'_{12}-\mathrm{readLen})!}{\mathrm{expr}'_{12}!(\mathrm{tranLen}'_{12}-\mathrm{readLen})!}} \\
&= \binom{\mathrm{expr}'_{12}}{\mathrm{expr}_1} \frac{\prod_{i\in\{1,2\}} \frac{(\mathrm{expr}_i+\mathrm{tranLen}_i-\mathrm{readLen})!}{(\mathrm{tranLen}_i-\mathrm{readLen})!}}{\frac{(\mathrm{expr}'_{12}+\mathrm{tranLen}'_{12}-\mathrm{readLen})!}{(\mathrm{tranLen}'_{12}-\mathrm{readLen})!}} \\
&= \binom{\mathrm{expr}'_{12}}{\mathrm{expr}_1} \frac{\prod_{i\in\{1,2\}} (\mathrm{tranLen}_i-\mathrm{readLen}+1)^{\overline{(\mathrm{expr}_i)}}}{(\mathrm{tranLen}'_{12}-\mathrm{readLen}+1)^{\overline{(\mathrm{expr}'_{12})}}}
\end{aligned}
$$

since $\mathrm{expr}'_{12} = \mathrm{expr}_1 + \mathrm{expr}_2$. Here $(n)^{\overline{(k)}} = n\cdot(n+1)\cdots(n+k-1) = (n+k-1)!/(n-1)!$, so that

$$
\begin{aligned}
&\frac{(\mathrm{expr}_i+\mathrm{tranLen}_i-\mathrm{readLen})!}{(\mathrm{tranLen}_i-\mathrm{readLen})!} \\
&= \frac{((\mathrm{tranLen}_i-\mathrm{readLen}+1)+(\mathrm{expr}_i)-1)!}{((\mathrm{tranLen}_i-\mathrm{readLen}+1)-1)!} \\
&= (\mathrm{tranLen}_i-\mathrm{readLen}+1)^{\overline{(\mathrm{expr}_i)}}
\end{aligned}
$$

Thus, in summary:

$$
\begin{aligned}
&\frac{P(\mathrm{contig},\mathrm{coverage},\mathrm{contigsFromTrans}'|\mathrm{trans}',\mathrm{expr}')}{P(\mathrm{contig},\mathrm{coverage},\mathrm{contigsFromTrans}|\mathrm{trans},\mathrm{expr})} \\
&= \frac{\binom{\mathrm{numContigsFromTran}'_{12}}{\mathrm{numContigsFromTran}_1}}{\prod_c \prod_l \binom{\mathrm{numContigsWithStats}'_{12,c,l}}{\mathrm{numContigsWithStats}_{1,c,l}}} \frac{\frac{(\mathrm{tranLen}'_{12}+1-\sum_{j\in\mathrm{contigsFromTran}'_{12}} \mathrm{contigLen}_j)_{\underline{(\mathrm{numContigsFromTran}'_{12})}}}{\prod_{i\in\{1,2\}} (\mathrm{tranLen}_i+1-\sum_{j\in\mathrm{contigsFromTran}_i} \mathrm{contigLen}_j)_{\underline{(\mathrm{numContigsFromTran}_i)}}}}{\binom{\mathrm{numContigsFromTran}'_{12}}{\mathrm{numContigsFromTran}_1}} \\
&\quad \cdot \binom{\mathrm{expr}'_{12}}{\mathrm{expr}_1} \frac{\prod_{i\in\{1,2\}} (\mathrm{tranLen}_i-\mathrm{readLen}+1)^{\overline{(\mathrm{expr}_i)}}}{(\mathrm{tranLen}'_{12}-\mathrm{readLen}+1)^{\overline{(\mathrm{expr}'_{12})}}} \\
&= \frac{\frac{(\mathrm{tranLen}'_{12}+1-\sum_{j\in\mathrm{contigsFromTran}'_{12}} \mathrm{contigLen}_j)_{\underline{(\mathrm{numContigsFromTran}'_{12})}}}{\prod_{i\in\{1,2\}} (\mathrm{tranLen}_i+1-\sum_{j\in\mathrm{contigsFromTran}_i} \mathrm{contigLen}_j)_{\underline{(\mathrm{numContigsFromTran}_i)}}}}{\prod_c \prod_l \binom{\mathrm{numContigsWithStats}'_{12,c,l}}{\mathrm{numContigsWithStats}_{1,c,l}}} \binom{\mathrm{expr}'_{12}}{\mathrm{expr}_1} \frac{\prod_{i\in\{1,2\}} (\mathrm{tranLen}_i-\mathrm{readLen}+1)^{\overline{(\mathrm{expr}_i)}}}{(\mathrm{tranLen}'_{12}-\mathrm{readLen}+1)^{\overline{(\mathrm{expr}'_{12})}}}
\end{aligned}
$$

The log odds is:

$$\begin{aligned}
&\log P(\text{contig}, \text{coverage}, \text{contigsFromTrans}' | \text{trans}', \text{expr}') - \log P(\text{contig}, \text{coverage}, \text{contigsFromTrans} | \text{trans}, \text{expr}) \\
&= \log\left((\text{tranLen}'_{12} + 1 - \sum_{j \in \text{contigsFromTran}'_{12}} \text{contigLen}_j)_{\underline{(\text{numContigsFromTran}'_{12})}} \right) \\
&\quad - \left(\sum_{i \in \{1,2\}} \log\left((\text{tranLen}_i + 1 - \sum_{j \in \text{contigsFromTran}_i} \text{contigLen}_j)_{\underline{(\text{numContigsFromTran}_i)}} \right) \right) \\
&\quad - \left(\sum_c \sum_l \log \binom{\text{numContigsWithStats}'_{12,c,l}}{\text{numContigsWithStats}_{1,c,l}} \right) \\
&\quad + \log \binom{\text{expr}'_{12}}{\text{expr}_1} \\
&\quad + \left(\sum_{i \in \{1,2\}} \log\left((\text{tranLen}_i - \text{readLen} + 1)^{\overline{(\text{expr}_i)}} \right) \right) \\
&\quad - \log\left((\text{tranLen}'_{12} - \text{readLen} + 1)^{\overline{(\text{expr}'_{12})}} \right)
\end{aligned}$$

4.1.4 Summary of the log odds

In summary,

$$
\begin{aligned}
&\log P(\mathrm{contig}, \mathrm{coverage}, \mathrm{contigsFromTrans}, \mathrm{trans}, \mathrm{expr}, \mathrm{reads}) - \log P(\mathrm{contig}, \mathrm{coverage}, \mathrm{contigsFromTrans}', \mathrm{trans}', \mathrm{expr}', \mathrm{reads}) \\
&= \left(\mathrm{expr}'_{12} \log \mathrm{expr}'_{12} - \sum_{j\in\{1,2\}} \mathrm{expr}_j \log \mathrm{expr}_j\right) \\
&\quad - \left(\mathrm{expr}'_{12} \log(\mathrm{tranLen}'_{12} - \mathrm{readLen} + 1) - \sum_{j\in\{1,2\}} \mathrm{expr}_j \log(\mathrm{tranLen}_j - \mathrm{readLen} + 1)\right) \\
&\quad + \frac{1}{2} \log \mathrm{numReads} \\
&\quad - \log\left(\frac{\mu_0(1-p_0)^{r_0}}{\mathrm{numReads} \cdot \mathrm{numTrans}}\right) \\
&\quad + \left(\log\binom{\mathrm{tranLen}'_{12} + r_0 - 1}{\mathrm{tranLen}'_{12}} - \sum_{i\in\{1,2\}} \log\binom{\mathrm{tranLen}_i + r_0 - 1}{\mathrm{tranLen}_i}\right) \\
&\quad + \left(\mathrm{tranLen}'_{12} - \sum_{i\in\{1,2\}} \mathrm{tranLen}_i\right) \log(p_0/4) \\
&\quad + \log\left(\left(\mathrm{tranLen}'_{12} + 1 - \sum_{j\in\mathrm{contigsFromTran}'_{12}} \mathrm{contigLen}_j\right)_{\underline{(\mathrm{numContigsFromTran}'_{12})}} \right) \\
&\quad - \left(\sum_{i\in\{1,2\}} \log\left(\left(\mathrm{tranLen}_i + 1 - \sum_{j\in\mathrm{contigsFromTran}_i} \mathrm{contigLen}_j\right)_{\underline{(\mathrm{numContigsFromTran}_i)}} \right) \right) \\
&\quad - \left(\sum_c \sum_l \log\binom{\mathrm{numContigsWithStats}'_{12,c,l}}{\mathrm{numContigsWithStats}_{1,c,l}}\right) \\
&\quad + \log\binom{\mathrm{expr}'_{12}}{\mathrm{expr}_1} \\
&\quad + \left(\sum_{i\in\{1,2\}} \log\left((\mathrm{tranLen}_i - \mathrm{readLen} + 1)^{\overline{(\mathrm{expr}_i)}} \right) \right) \\
&\quad - \log\left((\mathrm{tranLen}'_{12} - \mathrm{readLen} + 1)^{\overline{(\mathrm{expr}'_{12})}} \right)
\end{aligned}
$$

4.2 How should one choose $\mathrm{tranLen}'$?

We use the same notation from the previous subsection.

Claim: If one joins transcripts 1 and 2 into a single transcript 12, then J is maximized by setting $\mathrm{tranLen}'_{12} = \mathrm{tranLen}_1 + \mathrm{tranLen}_2$, i.e., by setting $\mathrm{tranLen}'_{12}$ as small as possible.

Proof: We look at all the terms of the log odds $\log P(\mathrm{contig}, \mathrm{coverage}, \mathrm{contigsFromTrans}', \mathrm{trans}', \mathrm{expr}', \mathrm{reads}) - \log P(\mathrm{contig}, \mathrm{coverage}, \mathrm{co}$ that include $\mathrm{tranLen}'_{12}$. The terms that include $\mathrm{tranLen}'_{12}$ are:

1. First, the term $-\mathrm{expr}'_{12} \log(\mathrm{tranLen}'_{12} - \mathrm{readLen} + 1)$ is contributed by $\log P(\mathrm{reads}|\mathrm{trans}', \mathrm{expr}') - \log P(\mathrm{reads}|\mathrm{trans}, \mathrm{expr})$.
2. Second, the terms

(i) $\log\binom{\text{tranLen}'_{12}+r_0-1}{\text{tranLen}'_{12}}+\log((r_0-1)!)$ and

(ii) $\text{tranLen}'_{12}\log(p_0/4)$

are contributed by $\log P(\text{trans}',\text{expr}')-\log P(\text{trans},\text{expr})$. (We added the $\log((r_0-1)!)$ term because part of the log binomial coefficient this is constant wrt $\text{tranLen}'_{12}$.)

3. Third, the terms

(i) $\log\left(\text{tranLen}'_{12}+1-\sum_{j\in\text{contigsFromTran}'_{12}}\text{contigLen}_j\right)_{\underline{(\text{numContigsFromTran}'_{12})}}$ and

(ii) $-\log\left(\text{tranLen}'_{12}-\text{readLen}+1\right)^{\overline{(\text{expr}'_{12})}}$

are contributed by $\log P(\text{contig},\text{coverage},\text{contigsFromTrans}'|\text{trans}',\text{expr}')-\log P(\text{contig},\text{coverage},\text{contigsFromTrans}|\text{trans},\text{expr})$

We want to find the $\text{tranLen}'_{12}$ that maximizes the sum of these terms. To make this easier, we will treat the $\text{tranLen}'_{12}$ as a real number and take derivatives. We will also use the following shorthand:

1. $t=\text{tranLen}'_{12}$.
2. $x=\text{expr}'_{12}$.
3. $a=\text{readLen}-1$.
4. $b=r_0-1$.
5. $c=\log(p_0/4)$.
6. $d=1-\sum_{j\in\text{contigsFromTran}'_{12}}\text{contigLen}_j$.
7. $h=\text{numContigsFromTran}'_{12}$.

(Note: t,x,a,b,h are positive, while c and d are negative. Also we require that $t>(1-d)+1=d$.) The objective is:

$$J(\text{tranLen}'_{12})=-\text{expr}'_{12}\log(\text{tranLen}'_{12}-\text{readLen}+1)+\log\binom{\text{tranLen}'_{12}+r_0-1}{\text{tranLen}'_{12}}+\log((r_0-1)!)+\text{tranLen}'_{12}\log(p_0/4)$$
$$+\log\left(\text{tranLen}'_{12}+1-\sum_{j\in\text{contigsFromTran}'_{12}}\text{contigLen}_j\right)_{\underline{(\text{numContigsFromTran}'_{12})}}-\log\left(\text{tranLen}'_{12}-\text{readLen}+1\right)^{\overline{(\text{expr}'_{12})}}$$

i.e.,

$$J(t)=-x\log(t-a)+\log\binom{t+b}{t}+\log(b!)+tc+\log\left(t+d\right)_{\underline{(h)}}-\log\left(t-a\right)^{\overline{(x)}}$$

Note:

1. $\log\binom{t+b}{t}+\log(b!)=\log((t+b)!)-\log(t!)-\log(b!)+\log(b!)=\log((t+b)!)-\log(t!)=\log[1\cdot2\cdots(t+b)]-\log[1\cdot2\cdots t]=\sum_{i=1}^{t+b}\log(i)-\sum_{i=1}^{t}\log(i)=\sum_{i=t+1}^{t+b}\log(i)=\sum_{i=1}^{b}\log(t+i)$.
2. $\log\left(t+d\right)_{\underline{(h)}}=\log[(t+d)!/(t+d-h)!]=\log[(t+d-h+1)\cdot(t+d-h+2)\cdots(t+d)]=\sum_{i=1}^{h}\log(t+d-h+i)$.
3. $\log\left(t-a\right)^{\overline{(x)}}=\log[(t-a)\cdot(t-a+1)\cdots(t-a+x-1)]=\sum_{i=0}^{x-1}\log(t-a+i)=\sum_{i=1}^{x}\log(t-a+1+i)$.

So:

$$J(t)=-x\log(t-a)+\left(\sum_{i=1}^{b}\log(t+i)\right)+tc+\left(\sum_{i=1}^{h}\log(t+d-h+i)\right)-\left(\sum_{i=1}^{x}\log(t-a+1+i)\right)$$

And the derivative:

$$J'(t)=-\frac{x}{t-a}+\left(\sum_{i=1}^{b}\frac{1}{t+i}\right)+c+\left(\sum_{i=1}^{h}\frac{1}{t+d-h+i}\right)-\left(\sum_{i=1}^{x}\frac{1}{t-a+1+i}\right)$$

4.3 How to choose which transcripts to merge?

For the default Trinity assembly on the real Trinity mouse data, there are about 50000 contigs. There are about $50000^2/2 = 1,250,000,000$ pairs of contigs. Thus any approach that requires us to compute something for all pairs is a nonstarter. This rules out looking at all pairs of contigs and deciding which ones to merge based on some pairwise criterion.

Instead, we will (i) build an ordered list of contigs and (ii) decide which ones to merge based on the list. This requires us to find some bounds. I.e., we want to be able to say that if two contigs differ more than a certain amount according to the list's ordering criterion, then combining them will be a bad idea.

References

Nour-Eddine Fahssi. “Polynomial Triangles Revisited”, 2012. http://arxiv.org/abs/1202.0228.

K. Balasubramanian, R. Viveros and N. Balakrishnan, “Some discrete distributions related to extended Pascal triangles”, Fibonacci Quarterly, 33 (1995), 415-425. http://www.fq.math.ca/Scanned/33-5/balasubramanian.pdf

Eric S. Lander, Michael S. Waterman. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*, Volume 2, Issue 3, April 1988. http://www.sciencedirect.com/science/article/pii/0888754388900079

Michael C. Wendl and W. Brad Barbazuk. Extension of Lander-Waterman theory for sequencing filtered DNA libraries. *BMC Bioinformatics* 2005, 6:245.

Yee Whye Teh, Dirichlet Process, 2007. http://www.gatsby.ucl.ac.uk/~ywteh/research/npbayes/dp.pdf.

Rick Durrett, *Probability Models for DNA Sequence Evolution*, Second Edition, 2008. www.math.cornell.edu/~durrett/Gbook/PM4DNA_0317.pdf

Fang Xu, The Poisson-Dirichlet distribution and Ewens sampling formula, November 9, 2010. http://www.math.mcmaster.ca/~volterwd/slidepaper/FangXu.pdf

J. Pitman, Exchangeable and partially exchangeable random partitions, *Probability Theory and Related Fields* 102, 145–158.

Hajime Yamato, Sibuya Masaaki, and Toshifumi Nomachi, Ordered sample from two-parameter GEM distribution, *Statistics and Probability Letters*, 2001.

EXTRA STUFF

4.3.1 How does $(\mathrm{Contigs}, \mathrm{Coverage}, \mathrm{ContigsFromTrans}) | \{\mathrm{Reads}, \mathrm{Trans}, \mathrm{Expr}\}$ change when we join two transcripts together?

Recall that

$$\begin{aligned}
&P(\mathrm{contigs}, \mathrm{coverage}, \mathrm{contigsFromTrans} | \mathrm{reads}, \mathrm{trans}, \mathrm{expr}) \\
&= P(\mathrm{numContigsFromTran} | \mathrm{trans}, \mathrm{expr}) \\
&\quad \cdot P(\mathrm{contigsFromTran} | \mathrm{numContigsFromTran}, \mathrm{trans}, \mathrm{expr}) \\
&\quad \cdot P(\mathrm{coverage} | \mathrm{contigsFromTran}, \mathrm{numContigsFromTran}, \mathrm{trans}, \mathrm{expr}) \\
&\quad \cdot P(\mathrm{contigLen} | \mathrm{coverage}, \mathrm{contigsFromTrans}, \mathrm{trans}, \mathrm{expr}) \\
&\quad \cdot P(\mathrm{contig} | \mathrm{contigLen}, \mathrm{coverage}, \mathrm{contigsFromTrans}, \mathrm{trans}, \mathrm{expr})
\end{aligned}$$

We will address each of these separately.

4.3.1.1 How does $P(\mathrm{numContigsFromTran} | \mathrm{trans}, \mathrm{expr})$ change when we join two transcripts together?

Note

$$\begin{aligned}
&\log P(\mathrm{numContigsFromTran} | \mathrm{trans}, \mathrm{expr}) - \log P(\mathrm{numContigsFromTran}' | \mathrm{trans}', \mathrm{expr}') \\
&= \left[\sum_{i=1}^{\mathrm{numTrans}} \log \mathrm{Binomial}(\mathrm{numContigsFromTran}_i | \mathrm{tranLen}_i, \alpha_i (1 - \alpha_i)^{\mathrm{readLen}-\mathrm{ws}}) \right] \\
&\quad - \left[\sum_{i=1}^{\mathrm{numTrans}'} \log \mathrm{Binomial}(\mathrm{numContigsFromTran}'_i | \mathrm{tranLen}'_i, \alpha'_i (1 - \alpha'_i)^{\mathrm{readLen}-\mathrm{ws}}) \right] \\
&= \log \mathrm{Binomial}(\mathrm{numContigsFromTran}_1 | \mathrm{tranLen}_1, \alpha_1 (1 - \alpha_1)^{\mathrm{readLen}-\mathrm{ws}}) \\
&\quad + \log \mathrm{Binomial}(\mathrm{numContigsFromTran}_2 | \mathrm{tranLen}_2, \alpha_2 (1 - \alpha_2)^{\mathrm{readLen}-\mathrm{ws}}) \\
&\quad - \log \mathrm{Binomial}(\mathrm{numContigsFromTran}'_{12} | \mathrm{tranLen}'_{12}, \alpha'_{12} (1 - \alpha'_{12})^{\mathrm{readLen}-\mathrm{ws}})
\end{aligned}$$

where $\alpha_i = 1 - (1 - \mathrm{tranLen}_i^{-1})^{\mathrm{expr}_i}$.

4.3.1.2 How does $P(\text{contigsFromTran}|\text{numContigsFromTran},\text{trans},\text{expr})$ change when we join two transcripts together?

Note

$$\begin{aligned}
&\log P(\text{contigsFromTran}|\text{numContigsFromTran},\text{trans},\text{expr})\\
&-\log P(\text{contigsFromTran}'|\text{numContigsFromTran}',\text{trans}',\text{expr}')\\
&=-\log\binom{\text{numContigs}}{\text{numContigsFromTran}_i : i=1,\ldots,\text{numTrans}}\\
&\quad+\log\binom{\text{numContigs}}{\text{numContigsFromTran}'_i : i=1,\ldots,\text{numTrans}'}\\
&=-\log(\text{numContigs}!)+\sum_{i=1}^{\text{numTrans}}\log(\text{numContigsFromTran}_i!)\\
&\quad+\log(\text{numContigs}!)-\sum_{i=1}^{\text{numTrans}'}\log(\text{numContigsFromTran}'_i!)\\
&=\log(\text{numContigsFromTran}_1!)+\log(\text{numContigsFromTran}_2!)-\log(\text{numContigsFromTran}'_{12}!)\\
&=\log(\text{numContigsFromTran}_1!)+\log(\text{numContigsFromTran}_2!)-\log((\text{numContigsFromTran}_1+\text{numContigsFromTran}_2)!)\\
&=-\log\binom{\text{numContigsFromTran}_1+\text{numContigsFromTran}_2}{\text{numContigsFromTran}_1}
\end{aligned}$$

4.3.1.3 How does $P(\text{coverage}|\text{contigsFromTran},\text{numContigsFromTran},\text{trans},\text{expr})$ change when we join two transcripts together?

Note

$$\begin{aligned}
&\log P(\text{coverage}|\text{contigsFromTran},\text{numContigsFromTran},\text{trans},\text{expr})\\
&\quad-\log P(\text{coverage}|\text{contigsFromTran}',\text{numContigsFromTran}',\text{trans}',\text{expr}')\\
&=\log\left(\prod_{i=1}^{\text{numTrans}}\frac{\prod_{j\in\text{contigsFromTran}_i}\text{Geometric}(\text{coverage}_j|\gamma_i)}{\text{NegativeBinomial}(\text{expr}_i|\text{numContigsFromTran}_i,1-\gamma_i)}\right)\\
&\quad-\log\left(\prod_{i=1}^{\text{numTrans}'}\frac{\prod_{j\in\text{contigsFromTran}'_i}\text{Geometric}(\text{coverage}'_j|\gamma'_i)}{\text{NegativeBinomial}(\text{expr}'_i|\text{numContigsFromTran}'_i,1-\gamma'_i)}\right)\\
&=\left[\sum_{i=1}^{\text{numTrans}}\Big(\sum_{j\in\text{contigsFromTran}_i}\log\text{Geometric}(\text{coverage}_j|\gamma_i)\Big)-\log\text{NegativeBinomial}(\text{expr}_i|\text{numContigsFromTran}_i,1-\gamma_i)\right]\\
&\quad-\left[\sum_{i=1}^{\text{numTrans}'}\Big(\sum_{j\in\text{contigsFromTran}'_i}\log\text{Geometric}(\text{coverage}_j|\gamma'_i)\Big)-\log\text{NegativeBinomial}(\text{expr}'_i|\text{numContigsFromTran}'_i,1-\gamma'_i)\right]\\
&=\Big(\sum_{j\in\text{contigsFromTran}_1}\log\text{Geometric}(\text{coverage}_j|\gamma_1)\Big)-\log\text{NegativeBinomial}(\text{expr}_1|\text{numContigsFromTran}_1,1-\gamma_1)\\
&\quad+\Big(\sum_{j\in\text{contigsFromTran}_2}\log\text{Geometric}(\text{coverage}_j|\gamma_2)\Big)-\log\text{NegativeBinomial}(\text{expr}_2|\text{numContigsFromTran}_2,1-\gamma_2)\\
&\quad-\Big(\sum_{j\in\text{contigsFromTran}'_{12}}\log\text{Geometric}(\text{coverage}_j|\gamma'_{12})\Big)+\log\text{NegativeBinomial}(\text{expr}'_{12}|\text{numContigsFromTran}'_{12},1-\gamma'_{12})\\
&=-\log\text{NegativeBinomial}(\text{expr}_1|\text{numContigsFromTran}_1,1-\gamma_1)\\
&\quad-\log\text{NegativeBinomial}(\text{expr}_2|\text{numContigsFromTran}_2,1-\gamma_2)\\
&\quad+\log\text{NegativeBinomial}(\text{expr}'_{12}|\text{numContigsFromTran}'_{12},1-\gamma'_{12})
\end{aligned}$$

where $\gamma_i=(1-\alpha_i)^{\text{readLen}-\text{ws}}$. The last equality holds because $\text{contigsFromTran}'_{12}=\text{contigsFromTran}_1\cup\text{contigsFromTran}_2$.

4.3.1.4 How does $P(\mathrm{contigLen}|\mathrm{coverage},\mathrm{contigsFromTrans},\mathrm{trans},\mathrm{expr})$ change when we join two transcripts together?

Note

$$\begin{aligned}
&\log P(\mathrm{contigLen}|\mathrm{coverage},\mathrm{contigsFromTrans},\mathrm{trans},\mathrm{expr})\\
&\quad-\log P(\mathrm{contigLen}|\mathrm{coverage},\mathrm{contigsFromTrans}',\mathrm{trans}',\mathrm{expr}')\\
&=\log\left(\prod_{i=1}^{\mathrm{numTrans}}\frac{\prod_{j\in\mathrm{contigsFromTran}_i}\binom{\mathrm{coverage}_j-1}{\mathrm{contigLen}_j-\mathrm{readLen}}_{\mathrm{readLen}+1}}{\sum_{L=0}^{\mathrm{tranLen}_i-\mathrm{numContigsFromTran}_i\cdot\mathrm{readLen}}\binom{\mathrm{expr}_i-\mathrm{numContigsFromTran}_i}{L}_{\mathrm{readLen}+1}}\right)\\
&\quad-\log\left(\prod_{i=1}^{\mathrm{numTrans}'}\frac{\prod_{j\in\mathrm{contigsFromTran}'_i}\binom{\mathrm{coverage}_j-1}{\mathrm{contigLen}_j-\mathrm{readLen}}_{\mathrm{readLen}+1}}{\sum_{L=0}^{\mathrm{tranLen}'_i-\mathrm{numContigsFromTran}'_i\cdot\mathrm{readLen}}\binom{\mathrm{expr}'_i-\mathrm{numContigsFromTran}'_i}{L}_{\mathrm{readLen}+1}}\right)\\
&=\log\left(\frac{\prod_{j\in\mathrm{contigsFromTran}_1}\binom{\mathrm{coverage}_j-1}{\mathrm{contigLen}_j-\mathrm{readLen}}_{\mathrm{readLen}+1}}{\sum_{L=0}^{\mathrm{tranLen}_1-\mathrm{numContigsFromTran}_1\cdot\mathrm{readLen}}\binom{\mathrm{expr}_1-\mathrm{numContigsFromTran}_1}{L}_{\mathrm{readLen}+1}}\right)\\
&\quad+\log\left(\frac{\prod_{j\in\mathrm{contigsFromTran}_2}\binom{\mathrm{coverage}_j-1}{\mathrm{contigLen}_j-\mathrm{readLen}}_{\mathrm{readLen}+1}}{\sum_{L=0}^{\mathrm{tranLen}_2-\mathrm{numContigsFromTran}_2\cdot\mathrm{readLen}}\binom{\mathrm{expr}_2-\mathrm{numContigsFromTran}_2}{L}_{\mathrm{readLen}+1}}\right)\\
&\quad-\log\left(\frac{\prod_{j\in\mathrm{contigsFromTran}'_{12}}\binom{\mathrm{coverage}_j-1}{\mathrm{contigLen}_j-\mathrm{readLen}}_{\mathrm{readLen}+1}}{\sum_{L=0}^{\mathrm{tranLen}'_{12}-\mathrm{numContigsFromTran}'_{12}\cdot\mathrm{readLen}}\binom{\mathrm{expr}'_{12}-\mathrm{numContigsFromTran}'_{12}}{L}_{\mathrm{readLen}+1}}\right)\\
&=-\log\left(\sum_{L=0}^{\mathrm{tranLen}_1-\mathrm{numContigsFromTran}_1\cdot\mathrm{readLen}}\binom{\mathrm{expr}_1-\mathrm{numContigsFromTran}_1}{L}_{\mathrm{readLen}+1}\right)\\
&\quad-\log\left(\sum_{L=0}^{\mathrm{tranLen}_2-\mathrm{numContigsFromTran}_2\cdot\mathrm{readLen}}\binom{\mathrm{expr}_2-\mathrm{numContigsFromTran}_2}{L}_{\mathrm{readLen}+1}\right)\\
&\quad+\log\left(\sum_{L=0}^{\mathrm{tranLen}'_{12}-\mathrm{numContigsFromTran}'_{12}\cdot\mathrm{readLen}}\binom{\mathrm{expr}'_{12}-\mathrm{numContigsFromTran}'_{12}}{L}_{\mathrm{readLen}+1}\right)
\end{aligned}$$

If $\mathrm{tranLen}_i>\mathrm{readLen}\cdot\mathrm{expr}_i$, then the denominators above are $(\mathrm{readLen}+1)^{\mathrm{expr}_i-\mathrm{numContigsFromTran}_i}$, in which case,

$$\begin{aligned}
&\log P(\mathrm{contigLen}|\mathrm{coverage},\mathrm{contigsFromTrans},\mathrm{trans},\mathrm{expr})\\
&\quad-\log P(\mathrm{contigLen}|\mathrm{coverage},\mathrm{contigsFromTrans}',\mathrm{trans}',\mathrm{expr}')\\
&=-\log\left((\mathrm{readLen}+1)^{\mathrm{expr}_1-\mathrm{numContigsFromTran}_1}\right)\\
&\quad-\log\left((\mathrm{readLen}+1)^{\mathrm{expr}_2-\mathrm{numContigsFromTran}_2}\right)\\
&\quad+\log\left((\mathrm{readLen}+1)^{\mathrm{expr}'_{12}-\mathrm{numContigsFromTran}'_{12}}\right)\\
&=-(\mathrm{expr}_1-\mathrm{numContigsFromTran}_1)\log(\mathrm{readLen}+1)\\
&\quad-(\mathrm{expr}_2-\mathrm{numContigsFromTran}_2)\log(\mathrm{readLen}+1)\\
&\quad+(\mathrm{expr}'_{12}-\mathrm{numContigsFromTran}'_{12})\log(\mathrm{readLen}+1)\\
&=0
\end{aligned}$$

4.3.1.5 How does $P(\text{contig}|\text{contigLen},\text{coverage},\text{contigsFromTrans},\text{trans},\text{expr})$ change when we join two transcripts together?

Note

$$\begin{aligned}
&\log P(\text{contig}|\text{contigLen},\text{coverage},\text{contigsFromTrans},\text{trans},\text{expr})\\
&-\log P(\text{contig}|\text{contigLen},\text{coverage},\text{contigsFromTrans}',\text{trans}',\text{expr}')\\
&=\log\left(\prod_{i=1}^{\text{numTrans}}\binom{\text{numContigsFromTran}_i+\text{tranLen}_i-\sum_{j=1}^{\text{numContigsFromTran}_i}\text{contigLen}_j}{\text{numContigsFromTran}_i}^{-1}\right)\\
&\quad-\log\left(\prod_{i=1}^{\text{numTrans}'}\binom{\text{numContigsFromTran}'_i+\text{tranLen}'_i-\sum_{j=1}^{\text{numContigsFromTran}'_i}\text{contigLen}_j}{\text{numContigsFromTran}'_i}^{-1}\right)\\
&=-\log\binom{\text{numContigsFromTran}_1+\text{tranLen}_1-\sum_{j=1}^{\text{numContigsFromTran}_1}\text{contigLen}_j}{\text{numContigsFromTran}_1}\\
&\quad-\log\binom{\text{numContigsFromTran}_2+\text{tranLen}_2-\sum_{j=1}^{\text{numContigsFromTran}_2}\text{contigLen}_j}{\text{numContigsFromTran}_2}\\
&\quad+\log\binom{\text{numContigsFromTran}'_{12}+\text{tranLen}'_{12}-\sum_{j=1}^{\text{numContigsFromTran}'_{12}}\text{contigLen}_j}{\text{numContigsFromTran}'_{12}}
\end{aligned}$$

4.4 Distribution of $(\text{Contigs}, \text{Coverage}, \text{ContigsFromTrans})\,|\,\{\text{Reads}, \text{Trans}, \text{Expr}\}$

For the distribution of contigs (and their coverage and originating transcripts), we use a simple Lander-Waterman–like approach (Lander and Waterman (1988)). (We could also use Bo's approach, which is perhaps in some ways more accurate; the approach here is mainly designed to make maximization easier later on.)

First, like in Bo's current approach, we assume that the contigs (and coverage) are conditionally independent of the reads, given the transcripts (and expression), i.e., in symbols:

$$(\text{Contigs}, \text{Coverage}, \text{ContigsFromTrans})\,|\,\{\text{Reads}, \text{Trans}, \text{Expr}\} =_D (\text{Contigs}, \text{Coverage}, \text{ContigsFromTrans})\,|\,\{\text{Trans}, \text{Expr}\}$$

Recall that numReads and readLen are not random; these quantities will be very important for the distributions below.

Next, we decompose the conditional distribution of $(\text{Contigs}, \text{Coverage}, \text{ContigsFromTrans})$ as follows:

$$\begin{aligned}
&P(\text{contigs}, \text{coverage}, \text{contigsFromTrans}|\text{trans}, \text{expr}) \\
&= P(\text{numContigsFromTran}|\text{trans}, \text{expr}) \\
&\quad \cdot P(\text{contigsFromTran}|\text{numContigsFromTran}, \text{trans}, \text{expr}) \\
&\quad \cdot P(\text{coverage}|\text{contigsFromTran}, \text{numContigsFromTran}, \text{trans}, \text{expr}) \\
&\quad \cdot P(\text{contigLen}|\text{coverage}, \text{contigsFromTrans}, \text{trans}, \text{expr}) \\
&\quad \cdot P(\text{contig}|\text{contigLen}, \text{coverage}, \text{contigsFromTrans}, \text{trans}, \text{expr})
\end{aligned}$$

4.4.1 Distribution of $\text{NumContigsFromTran}\,|\,\{\text{Trans}, \text{Expr}\}$

We assume that the number of contigs from transcript i is conditionally independent of the number of contigs from transcript j. So

$$P(\text{numContigsFromTran}|\text{trans}, \text{expr}) = \prod_{i=1}^{\text{numTrans}} P(\text{numContigsFromTran}_i|\text{tran}_i, \text{expr}_i)$$

Next, we find $P(\text{numContigsFromTran}_i|\text{tran}_i, \text{expr}_i)$. According to Lander-Waterman, if a transcript has length l and expression x, then the probability it generates n contigs is $\text{Binomial}(n|l, \alpha(1-\alpha)^{\text{readLen}-\text{ws}})$, where $\alpha = 1-(1-l^{-1})^x$ is the probability that at least one of the x reads starts at a particular position.

Reason: First:

1. l^{-1} is the probability that a particular read starts at a particular position, assuming a uniform RSPD, and assuming (falsely) that reads can start even at the end of the transcript (otherwise it would be $(l-\text{readLen}+1)^{-1}$). (See Figure 3.)
2. So $1-l^{-1}$ is the probability that a particular read does not start at a particular position.
3. So $(1-l^{-1})^x$ is the probability that none of the x reads starts at a particular position, assuming that the read start positions are independent.
4. So $\alpha = 1-(1-l^{-1})^x$ is the probability that at least one of the x reads starts at a particular position.

Second: The "beginning of the end" of a contig occurs at position k if (i) at least one read starts at position k, and (ii) no read starts at any of the subsequent $\text{readLen}-\text{ws}$ positions. (See Figure 4.) Call this event B_k. The probability of this event is $\alpha(1-\alpha)^{\text{readLen}-\text{ws}}$, assuming (again) independence between read start positions.

Third: The number of contigs that come from a transcript is equal to the number of "beginning of the end" events B_k that occur. If we (i) assume that the B_k are independent (so that, for example, "beginning of the end" events can

```
01234567890123456789
-----
 -----
        [...]
               ----- 15 = 20 - 5
               +++++
                +++++
                 +++++
                  +++++
                   +++++ 19 = 20 - 1
```

Figure 3: What is the probability that a particular read starts from a particular position? Here transcript length is 20 and read length is 5. Reads can start at positions 0, 1, ..., 15, i.e., at $16 = 20 - 5 + 1 = \text{tranLen} - \text{readLen} + 1$ distinct positions. Assuming uniform RSPD, the chance of starting at any of these positions is $1/(\text{tranLen} - \text{readLen} + 1)$. Assuing, as we do for simplicity, that reads can start even at positions 16, 17, ..., 19, i.e., at $20 = \text{tranLen}$ distinct positions, the chance of starting at any of these positions is $1/\text{tranLen}$.

```
01234567890123456789
  -----      Read occurs at position 2.
   .....     Read fails to occur at position 3.
    .....    Read fails to occur at position 4.
     .....   Read fails to occur at position 5.
```

Figure 4: Example of a contig ending. Here transcript length is 20, read length is 5, and window size is 2. The contig ends when there is one read occurrence, followed by $3 = 5 - 2 = \text{readLen} - \text{ws}$ failures to occur.

occur even right next to each other), and (ii) we ignore edge effects (so that "beginning of the end" events have equal probability even at the rightmost end of the transcript), then the probability of getting n contigs is $\text{Binomial}(n|l, \alpha(1-\alpha)^{\text{readLen}-\text{ws}})$. (Note: Assumption (i) is grossly incorrect. Assumption (ii) is perhaps not too bad if there are long poly(A) tails. Alternatively, we could use the more complicated formulas found by Wendl and Barbazuk (2005), or Bo's approach, to take into account the edge effects. But this makes maximization more complicated later on.)

In summary:

$$P(\text{numContigsFromTran}_i|\text{tran}_i, \text{expr}_i) = \text{Binomial}(\text{numContigsFromTran}_i|\text{tranLen}_i, \alpha_i(1-\alpha_i)^{\text{readLen}-\text{ws}})$$

where $\alpha_i = 1 - (1 - \text{tranLen}_i^{-1})^{\text{expr}_i}$ is the probability that at least one of the expr_i reads starts at a particular position in transcript i.

4.4.2 Distribution of $\text{ContigsFromTran}|\{\text{NumContigsFromTran}, \text{Trans}, \text{Expr}\}$

We assume that the chance of any particular assignment of contigs to transcripts, conditional on the number of contigs from each transcript, is uniform over the relevant set. The relevant set is the assignments of $\text{numContigsFromTran}_i$ contigs to transcript i, for each i, without replacement, where there are $\text{numContigs} = \sum_{i=1}^{\text{numTrans}} \text{numContigsFromTran}_i$ contigs total. (numContigs is a deterministic function of numContigsFromTran.) The size of this set is given by the multinomial coefficient. So:

$$P(\text{contigsFromTran}|\text{numContigsFromTran}, \text{trans}, \text{expr}) = \binom{\text{numContigs}}{\text{numContigsFromTran}_i : i = 1, \ldots, \text{numTrans}}^{-1}$$

4.4.3 Distribution of $\text{Coverage}|\{\text{ContigsFromTrans}, \text{Trans}, \text{Expr}\}$

First, we stipulate that the coverage of the contigs from transcript i is conditionally independent of both the other transcripts and of the coverage of the contigs from other transcripts, conditional on the transcript i. Using the notation $X_{(i)} = (X_j : j \in \text{ContigsFromTran}_i)$, this gives

$$\begin{aligned}&P(\text{coverage}|\text{contigsFromTrans}, \text{trans}, \text{expr})\\&= \prod_{i=1}^{\text{numTrans}} P(\text{coverage}_{(i)}|\text{contigsFromTran}_i, \text{tran}_i, \text{expr}_i)\end{aligned}$$

If we also, for the moment, assume that the coverages of contigs from transcript i are also conditionally independent, letting $\text{Coverage}'_{(i)}$ be the conditionally independent version of the random variable, then

$$\begin{aligned}&P(\text{coverage}'_{(i)}|\text{contigsFromTran}_i, \text{tran}_i, \text{expr}_i)\\&= \prod_{j\in\text{contigsFromTran}_i} P(\text{coverage}'_j|\text{contigsFromTran}_i, \text{tran}_i, \text{expr}_i)\end{aligned}$$

and we can use Lander-Waterman to find $P(\text{coverage}'_j|\text{contigsFromTran}_i, \text{tran}_i, \text{expr}_i)$.

According to Lander-Waterman,

$$P(\text{coverage}'_j|\text{contigsFromTran}_i, \text{tran}_i, \text{expr}_i) = \text{Geometric}(\text{coverage}'_j|(1-\alpha_i)^{\text{readLen}-\text{ws}}).$$

Reason: Assume that contig j has already started, so there are currently c reads covering the left prefix of contig j, with rightmost read starting at position k within the transcript. The contig ends, with total coverage c, if $\text{readLen} - \text{ws}$ reads do not occur, starting at position $k+1$; this happens with probability $\gamma_i = (1-\alpha_i)^{\text{readLen}-\text{ws}}$ (see previous paragraph). Thus, the chance of getting coverage c, for any c, is the chance that an event with probability γ_i (namely, the absence of $\text{readLen} - \text{ws}$ reads in a row) fails to occur c times. This is the geometric distribution with termination probability γ_i. (We use the variant which includes 0 in the support, because if the termination event occurs in the first trial, then read coverage is 0. It would be more accurate to exclude this possibility - that that would make the formulas slightly more complicated.)

Above, we temporarily assumed that the coverages of contigs from transcript i are conditionally independent, but this is certainly not true, because the sum of the coverages equals the expression, i.e., $\sum_{j\in\text{ContigsFromTran}_i} \text{Coverage}_i = \text{Expr}_i$. Therefore we instead assume that the coverages of contigs from transcript i are conditionally independent, conditional not just on the transcript and its expression but also on the event that the sum of the coverages equals the expression, i.e., the event $B_i := \{\omega : \sum_{j\in\text{ContigsFromTran}_i}\}$:

$$\begin{aligned}&P(\text{coverage}_{(i)}|\text{contigsFromTran}_i, \text{tran}_i, \text{expr}_i)\\&= P(\text{coverage}_{(i)}|B_i, \text{contigsFromTran}_i, \text{tran}_i, \text{expr}_i)P(B_i|\text{contigsFromTran}_i, \text{tran}_i, \text{expr}_i)\\&\quad + P(\text{coverage}_{(i)}|B_i^c, \text{contigsFromTran}_i, \text{tran}_i, \text{expr}_i)P(B_i^c|\text{contigsFromTran}_i, \text{tran}_i, \text{expr}_i)\\&= P(\text{coverage}_{(i)}|B_i, \text{contigsFromTran}_i, \text{tran}_i, \text{expr}_i)\end{aligned}$$

The last equality holds because the conditional probability of B_i is stipulated to be 1.

It remains to specify $P(\text{coverage}_{(i)}|B_i, \text{contigsFromTran}_i, \text{tran}_i, \text{expr}_i)$. We saw above that under the assumption of conditional independence of the coverages, each coverage is conditionally distributed according to $\text{Geometric}(\gamma_i)$. Conditioning on B_i restricts the support of $\text{coverage}_{(i)}$, but does not otherwise change the distribution. Fact: In general, if $X_i \sim \text{Geometric}(p)$ iid, then $\sum_{i=1}^n X_i \sim \text{NegativeBinomial}(n, 1-p)$. See the next paragraph for a proof. Applying this to our case,

$$\begin{aligned}&P(\text{coverage}_{(i)}|B_i,\text{contigsFromTran}_i,\text{tran}_i,\text{expr}_i)\\&=\frac{\prod_{j\in\text{contigsFromTran}_i}\text{Geometric}(\text{coverage}_j|\gamma_i)}{\text{NegativeBinomial}(\text{expr}_i|\text{numContigsFromTran}_i,1-\gamma_i)}\end{aligned}$$

Reason for the fact mentioned in the previous paragraph, that in general, if $X_i \sim \text{Geometric}(p)$ iid, then $\sum_{i=1}^n X_i \sim \text{NegativeBinomial}(n,1-p)$: If X is a random variable, then let the MGF $M_X(t)=Ee^{tX}$. In general, if $M_X=M_Y$ then $X=_D Y$. If X is $\text{Geometric}(p)$ and Y is $\text{NegativeBinomial}(n,1-p)$, then $M_X(t)=\frac{p}{1-(1-p)e^t}$ and $M_Y(t)=\left(\frac{p}{1-(1-p)e^t}\right)^n$. The MGF of $\sum_{i=1}^n X_i$ is:

$$M_{\sum_{i=1}^n X_i}(t)=Ee^{t\sum_{i=1}^n X_i}=E\prod_{i=1}^n e^{tX_i}=\prod_{i=1}^n Ee^{tX_i}=\prod_{i=1}^n M_X(t)=M_X(t)^n=M_Y(t)$$

so $\sum_{i=1}^n X_i =_D Y$ as claimed.

In summary:

$$\begin{aligned}&P(\text{coverage}|\text{contigsFromTrans},\text{trans},\text{expr})\\&=\prod_{i=1}^{\text{numTrans}}P(\text{coverage}_{(i)}|\text{contigsFromTran}_i,\text{tran}_i,\text{expr}_i)\\&=\prod_{i=1}^{\text{numTrans}}P(\text{coverage}_{(i)}|B_i,\text{contigsFromTran}_i,\text{tran}_i,\text{expr}_i)\\&=\prod_{i=1}^{\text{numTrans}}\frac{\prod_{j\in\text{contigsFromTran}_i}\text{Geometric}(\text{coverage}_j|\gamma_i)}{\text{NegativeBinomial}(\text{expr}_i|\text{numContigsFromTran}_i,1-\gamma_i)}\end{aligned}$$

4.4.4 Distribution of $\text{ContigLen}|\{\text{Coverage},\text{ContigsFromTrans},\text{Trans},\text{Expr}\}$

Just as for the coverage, we assume that the contig lengths for contigs from one transcript are conditionally independent of other transcripts and their contigs:

$$\begin{aligned}&P(\text{contigLen}|\text{coverage},\text{contigsFromTrans},\text{trans},\text{expr})\\&=\prod_{i=1}^{\text{numTrans}}P(\text{contigLen}_{(i)}|\text{coverage}_{(i)},\text{contigsFromTran}_i,\text{trans}_i,\text{expr}_i)\end{aligned}$$

Also, just as for the coverage, we (i) find how the contig lengths for a particular transcript would be distributed assuming conditional independence even among contigs from a single transcript, (ii) reinstatement of the dependence, by restricting the domain of the joint distribution and normalizing.

Let A_i be the event $\{\omega : \text{Coverage}_{(i)}=\text{coverage}_{(i)},\text{ContigsFromTran}_i=\text{contigsFromTran}_i,\text{Trans}_i=\text{trans}_i,\text{Expr}_i=\text{expr}_i\}$.

4.4.4.1 (i) The distribution, assuming conditional independence even among contigs from a single transcript

Let $\text{ContigLen}'_{(i)}$ be the conditionally independent version of $\text{ContigLen}_{(i)}$.

Recall (a) that we are assuming that all reads from a transcript start (i) independently of each other and (ii) uniformly along the length of the transcript, including even at the end of the transcript. I.e., the probability that a particular read

r from transcript i starts at position k of transcript i is tranLen_i^{-1}, independently of all other reads. Recall also (b) that, by definition, each contig is a contiguous collection of reads, in the sense that, if one enumerates the reads involved in this contig from left to right, there is always at least ws bases of overlap between each adjacent (in the enumeration) read. See Figure 5 for an example.

Consider a contig j from transcript i. Due to item (b), contig j's length $\text{contigLen}'_j$ must be between readLen and $\text{readLen} \cdot \text{coverage}_j$. Moreover, each possible way to arrange the coverage_j reads involved in contig j consistent with item (b) is equally likely, due to item (a). Therefore, the conditional probability of contig j having length $\text{contigLen}'_j$ is proportional to the number of ways the coverage_j reads can be arranged into a contig of length $\text{contigLen}'_j$. In symbols,

$$P(\text{contigLen}'_j|A_i) = \frac{\text{numArrangements}(\text{coverage}_j, \text{contigLen}'_j, \text{readLen})}{\sum_{l=\text{readLen}}^{\text{readLen}\cdot\text{coverage}_j} \text{numArrangements}(\text{coverage}_j, l, \text{readLen})}$$

Here numArrangements need not be considered to include different permutations of read identities, since doing so would just introduce the same factor $\text{coverage}_j!$ in both the numerator and denominator. (Cf. Figure 5.)

Claim: $\text{numArrangements}(\text{coverage}_j, \text{contigLen}'_j, \text{readLen}) = \binom{\text{coverage}_j - 1}{\text{contigLen}'_j - \text{readLen}}_{\text{readLen}+1}$, where

$$\binom{n}{k}_N = \begin{cases} 0 & \text{if } n < 0 \text{ or } k < 0 \text{ or } k > (N-1)\cdot n \\ 1 & \text{if } n = 0 \text{ and } k = 0 \\ \sum_{j=0}^{N-1} \binom{n-1}{k-j}_N & \text{otherwise} \end{cases}$$

Proof: Fix an enumeration of the coverage_j reads involved in contig j from left to right, as described in item (b) above. Each read can be thought to "newly contribute" a certain number of bases, in the sense that no previous (in the enumeration) read covers these bases, but the read in question does cover them. See Figure 6 for an example. The first (and leftmost) read always contributes readLen bases. Each remaining read can contribute between 0 bases (if it completely overlaps the previous read) and readLen bases (if it starts just after the previous read), assuming that we require window size of 0.

For each arrangement a of the reads, let $x = x(a)$ be the corresponding "new-contribution" vector, i.e., x_r is the number of bases newly contributed by read r in the arrangement. Note that there is a 1-1 correspondence between arrangements a and new-contribution vectors x. Thus the number of arrangements of the reads into a contig of length $\text{contigLen}'_j$ is equal to the number of valid new-contribution vectors, i.e., the number of vectors $x \in \{\text{readLen}\} \times \{0, \ldots, \text{readLen}\}^{\text{coverage}_j - 1}$ such that $\sum_{r=1}^{\text{coverage}_j - 1} x_r = \text{contigLen}'_j$.

How many such new-contribution vectors are there? The first read always newly covers readLen bases. Each new-contribution vector says how to allocate the remaining $\text{contigLen}'_j - \text{readLen}$ bases among the remaining $\text{coverage}_j - 1$ reads, allowing at most readLen positions to be allocated to each read. We can think of the bases as being unlabelled, in the sense that the new-contribution vectors do not keep track of *which* specific (identified) bases are newly covered by each read, but rather just *how many* bases are newly covered by each read. We can think of the reads as being labelled, in the sense that each new-contribution vector says how many bases are newly contributed by each read which is labelled by its position within the (fixed) enumeration. Thus, the number of new-contribution vectors is the number of distinct ways in which $k := \text{contigLen}'_j - \text{readLen}$ balls (i.e., bases) can be allocated to $n := \text{coverage}_j - 1$ urns (i.e., reads), allowing at most $N - 1 := \text{readLen}$ balls (i.e., bases) to fall in each urn (i.e., read). It is known that the number of ways to do this is $\binom{n}{k}_N$: see comments by N-E. Fahssi from OEIS entries `http://oeis.org/A008287` and `http://oeis.org/A035343`, and cf. also Fahssi (2012).

For empirical evidence of the claim, see `howManyWaysToCoverAContig.py`. ■

Claim: $\sum_{l=\text{readLen}}^{\text{readLen}\cdot\text{coverage}_j} \text{numArrangements}(\text{coverage}_j, l, \text{readLen}) = (\text{readLen}+1)^{\text{coverage}_j - 1}$.

```
0123456   contig
aaaaa
bbbbb
  ccccc

0123456   contig
aaaaa
 bbbbb
  ccccc

0123456   contig
aaaaa
  bbbbb
  ccccc
```

Figure 5: All the ways to get a contig of length 7 from 3 reads of length 5, when we do not consider read identity. If we consider read identity, then the number of ways is multiplied by 3!.

Proof: Note that

$$\begin{aligned}
&\sum_{l=\text{readLen}}^{\text{readLen}\cdot\text{coverage}_j} \text{numArrangements}(\text{coverage}_j, l, \text{readLen}) \\
&= \sum_{l=\text{readLen}}^{\text{readLen}\cdot\text{coverage}_j} \binom{\text{coverage}_j - 1}{l - \text{readLen}}_{\text{readLen}+1} \\
&= \sum_{l=0}^{\text{readLen}\cdot(\text{coverage}_j-1)} \binom{\text{coverage}_j - 1}{l}_{\text{readLen}+1}
\end{aligned}$$

the $(\text{coverage}_j - 1)$th row sum of the array with entries $(n,k) = \binom{n}{k}_{\text{readLen}+1}$. It is a fact (see `http://oeis.org/A027907` for the trinomial case) that these row sums are $n \mapsto N^n$, i.e., in our case, the $(\text{coverage}_j - 1)$th row sum is $(\text{readLen} + 1)^{\text{coverage}_j - 1}$.

For empirical evidence, see `howManyWaysToCoverAContig.py`. ■

Plugging in from the claims,

$$\begin{aligned}
P(\text{contigLen}'_j | A_i) &= \frac{\text{numArrangements}(\text{coverage}_j, \text{contigLen}'_j, \text{readLen})}{\sum_{l=\text{readLen}}^{\text{readLen}\cdot\text{coverage}_j} \text{numArrangements}(\text{coverage}_j, l, \text{readLen})} \\
&= \binom{\text{coverage}_j - 1}{\text{contigLen}'_j - \text{readLen}}_{\text{readLen}+1} (\text{readLen}+1)^{-(\text{coverage}_j - 1)}
\end{aligned}$$

(Note: This is essentially Eq. (2.4) in Fahssi (2012), with $k := \text{coverage}_j - 1$, $n := \text{contigLen}'_j - \text{readLen}$, $\mathbf{a} := (1,1,\ldots,1) \in \mathbb{N}^{\text{readLen}+1}$, and $N := \text{readLen} + 1$.)

(Miscellaneous note: The mean and variance of $\text{ContigLen}'_j | A_i$ are

$$\begin{aligned}
E(\text{ContigLen}'_j | A_i) &= \text{readLen} \cdot (\text{coverage}_j + 1)/2 \\
\text{Var}(\text{ContigLen}'_j | A_i) &= \text{readLen} \cdot (\text{readLen} + 2) \cdot (\text{coverage}_j - 1)/12
\end{aligned}$$

See `howManyWaysToCoverAContig.py`.)

4.4.4.2 (ii) The distribution, with the conditional dependence reinstated by restricting the support of the joint distribution and normalizing.

When we reinstate conditional dependence, we need to take into account the fact that

```
012345678901   contig
-----          contributes 5 new bases
   ..---       contributes 3 new bases
   .....       contributes 0 new bases
    ....-      contributes 1 new bases
       ..---   contributes 3 new bases
```

Figure 6: How many bases does each read contribute to a contig? Here contig length is 12 and read length is 5. Newly contributed bases in each read are marked by -, and other bases are marked by `.`. For example, read 1 starts at position 0 and read 2 starts at position 3, so read 2 contributes $3-0=3$ bases.

$\sum_{j\in\text{contigsFromTran}_i} \text{contigLen}_j \leq \text{tranLen}_i$, i.e., the support is smaller than when we assumed conditional independence. This does not affect the numerator of the pmf given in the last section, i.e., $\text{numArrangements}(\text{coverage}_j, \text{contigLen}'_j, \text{readLen})$ [assuming that the transcript is longer than the contig], since the number of arrangements was calculated without reference to the other contigs. In other words,

$$\begin{aligned}
P(\text{contigLen}_{(i)}|A_i) &= \frac{\prod_{j\in\text{contigsFromTran}_i} \text{numArrangements}(\text{coverage}_j, \text{contigLen}_j, \text{readLen})}{\text{denom}_i} \\
&= \frac{\prod_{j\in\text{contigsFromTran}_i} \binom{\text{coverage}_j - 1}{\text{contigLen}_j - \text{readLen}}_{\text{readLen}+1}}{\text{denom}_i}
\end{aligned}$$

for $\text{contigLen}_{(i)}$ such that $\sum_{j\in\text{contigsFromTran}_i} \text{contigLen}_j \leq \text{tranLen}_i$.

However, since the support is smaller, the reinstatement of conditional dependence does affect the denominator. Specifically, the denominator is

$$\begin{aligned}
\text{denom}_i &= \sum_{l:\sum_j l_j \leq \text{tranLen}_i} \prod_{j\in\text{contigsFromTran}_i} \text{numArrangements}(\text{coverage}_j, l_j, \text{readLen}) \\
&= \sum_{l:\sum_j l_j \leq \text{tranLen}_i} \prod_{j\in\text{contigsFromTran}_i} \binom{\text{coverage}_j - 1}{l_j - \text{readLen}}_{\text{readLen}+1}
\end{aligned}$$

where the $\sum_j$ is short for $\sum_{j\in\text{contigsFromTran}_i}$. This is the total number of ways to arrange coverage_{j_1} reads into a contig of length l_{j_1}, and coverage_{j_2} reads into a contig of length l_{j_2}, etc., for all contigs j_k from transcript i, subject to the total length of all the contigs from transcript i being less than or equal to the length of transcript i.

In order to evaluate denom_i, we use the following "Vandermonde convolution" identity from Fahssi (2012), Table 1:

$$\sum_{i+j=n} \binom{r}{i}_N \binom{s}{j}_N = \binom{r+s}{n}_N$$

Applying this to our case,

$$\begin{aligned}
&\sum_{l_{j_1}+l_{j_2}=L} \binom{\text{coverage}_{j_1} - 1}{l_{j_1} - \text{readLen}}_{\text{readLen}+1} \binom{\text{coverage}_{j_2} - 1}{l_{j_2} - \text{readLen}}_{\text{readLen}+1} \\
&= \sum_{l_{j_1}+l_{j_2}=L-2\cdot\text{readLen}} \binom{\text{coverage}_{j_1} - 1}{l_{j_1}}_{\text{readLen}+1} \binom{\text{coverage}_{j_2} - 1}{l_{j_2}}_{\text{readLen}+1} \\
&= \binom{\text{coverage}_{j_1} + \text{coverage}_{j_2} - 2}{L - 2\cdot\text{readLen}}_{\text{readLen}+1}
\end{aligned}$$

and by applying this recursively

$$\begin{aligned}
&\sum_{l:\sum_j l_j = L}\ \prod_{j\in\text{contigsFromTran}_i} \binom{\text{coverage}_j - 1}{l_j - \text{readLen}}_{\text{readLen}+1} \\
&= \binom{(\sum_j \text{coverage}_j) - \text{numContigsFromTran}_i}{L - \text{numContigsFromTran}_i \cdot \text{readLen}}_{\text{readLen}+1}
\end{aligned}$$

Thus, plugging in,

$$\begin{aligned}
\text{denom}_i &= \sum_{l:\sum_j l_j \le \text{tranLen}_i}\ \prod_{j\in\text{contigsFromTran}_i} \binom{\text{coverage}_j - 1}{l_j - \text{readLen}}_{\text{readLen}+1} \\
&= \sum_{L=0}^{\text{tranLen}_i}\ \sum_{l:\sum_j l_j = L}\ \prod_{j\in\text{contigsFromTran}_i} \binom{\text{coverage}_j - 1}{l_j - \text{readLen}}_{\text{readLen}+1} \\
&= \sum_{L=0}^{\text{tranLen}_i} \binom{(\sum_j \text{coverage}_j) - \text{numContigsFromTran}_i}{L - \text{numContigsFromTran}_i \cdot \text{readLen}}_{\text{readLen}+1} \\
&= \sum_{L=0}^{\text{tranLen}_i - \text{numContigsFromTran}_i \cdot \text{readLen}} \binom{(\sum_j \text{coverage}_j) - \text{numContigsFromTran}_i}{L}_{\text{readLen}+1} \\
&= \sum_{L=0}^{\text{tranLen}_i - \text{numContigsFromTran}_i \cdot \text{readLen}} \binom{\text{expr}_i - \text{numContigsFromTran}_i}{L}_{\text{readLen}+1}
\end{aligned}$$

a partial row sum. (See `howManyWaysToCoverAContig.py` for empirical verification of this formula.)

Recall from the definition of the polynomial coefficients $\binom{n}{k}_N$ that if $k > (N-1)\cdot n$, then the coefficient is 0. Thus, in the sum above, if $L > \text{readLen}\cdot(\text{expr}_i - \text{numContigsFromTran}_i)$, then the Lth term is 0. Therefore, if $\text{tranLen}_i > \text{readLen}\cdot\text{expr}_i$, then the partial row sum will be a full row sum, and $\text{denom}_i = (\text{readLen}+1)^{\text{expr}_i - \text{numContigsFromTran}_i}$.

Thus

$$\begin{aligned}
P(\text{contigLen}_{(i)}|A_i) &= \frac{\prod_{j\in\text{contigsFromTran}_i} \text{numArrangements}(\text{coverage}_j, \text{contigLen}_j, \text{readLen})}{\text{denom}_i} \\
&= \frac{\prod_{j\in\text{contigsFromTran}_i} \binom{\text{coverage}_j - 1}{\text{contigLen}_j - \text{readLen}}_{\text{readLen}+1}}{\sum_{L=0}^{\text{tranLen}_i - \text{numContigsFromTran}_i \cdot \text{readLen}} \binom{\text{expr}_i - \text{numContigsFromTran}_i}{L}_{\text{readLen}+1}}
\end{aligned}$$

for $\text{contigLen}_{(i)}$ such that $\sum_{j\in\text{contigsFromTran}_i} \text{contigLen}_j \le \text{tranLen}_i$.

In summary,

$$\begin{aligned}
&P(\text{contigLen}|\text{coverage}, \text{contigsFromTrans}, \text{trans}, \text{expr}) \\
&= \prod_{i=1}^{\text{numTrans}} P(\text{contigLen}_{(i)}|\text{coverage}_{(i)}, \text{contigsFromTran}_i, \text{trans}_i, \text{expr}_i) \\
&= \prod_{i=1}^{\text{numTrans}} \frac{\prod_{j\in\text{contigsFromTran}_i} \binom{\text{coverage}_j - 1}{\text{contigLen}_j - \text{readLen}}_{\text{readLen}+1}}{\sum_{L=0}^{\text{tranLen}_i - \text{numContigsFromTran}_i \cdot \text{readLen}} \binom{\text{expr}_i - \text{numContigsFromTran}_i}{L}_{\text{readLen}+1}}
\end{aligned}$$

4.4.5 Distribution of $\mathrm{Contig}|\{\mathrm{ContigLen}, \mathrm{Coverage}, \mathrm{NumContigsFromTran}_i, \mathrm{Tran}_i, \mathrm{Expr}_i)\}$

Again, we assume that the contigs from one transcript are conditionally independent of other transcripts and their contigs:

$$
\begin{aligned}
&P(\mathrm{contig}|\mathrm{contigLen}, \mathrm{coverage}, \mathrm{contigsFromTrans}, \mathrm{trans}, \mathrm{expr}) \\
&= \prod_{i=1}^{\mathrm{numTrans}} P(\mathrm{contig}_{(i)}|\mathrm{contigLen}_{(i)}, \mathrm{coverage}_{(i)}, \mathrm{contigsFromTran}_i, \mathrm{trans}_i, \mathrm{expr}_i)
\end{aligned}
$$

We stipulate that $P(\mathrm{contig}_{(i)}|\mathrm{contigLen}_{(i)}, \mathrm{coverage}_{(i)}, \mathrm{numContigsFromTran}_i, \mathrm{tran}_i, \mathrm{expr}_i)$ is uniform over the relevant set. The relevant set is defined by two constraints on the contigs: (i) each contig is a contiguous subsequence of its transcript, and (ii) all the contigs are disjoint subsequences. Note that we do not admit the possibility that contigs might be erroneous wrt the transcripts. So the relevant set is the set of possible contig placements subject to these constraints.

In each contig placement, there are $k = (\mathrm{tranLen}_i - \sum_{j=1}^{\mathrm{numContigsFromTran}_i} \mathrm{contigLen}_j)$ uncovered bases. For now, fix one ordering of the contigs. Each uncovered base can be placed before contig 1, after contig 1, after contig 2, ..., after contig $\mathrm{numContigsFromTran}_i$, i.e., at any of $n = (\mathrm{numContigsFromTran}_i + 1)$ relative positions. The number of ways of placing uncovered bases in such relative positions, where the different uncovered bases are treated as equivalent (so order doesn't matter), is equal to the number of multisets of k elements, each chosen with replacement from n choices, namely, $\binom{n+k-1}{k}$. See `numPositions.R` and Figure 7.

Therefore the conditional probability of the contig placements, and hence of the contigs, is

$$
\begin{aligned}
&P(\mathrm{contig}_{(i)}|\mathrm{contigLen}_{(i)}, \mathrm{coverage}_{(i)}, \mathrm{numContigsFromTran}_i, \mathrm{tran}_i, \mathrm{expr}_i) \\
&= \binom{n+k-1}{k}^{-1} = \binom{n+k-1}{n-1}^{-1} \\
&= \binom{\mathrm{numContigsFromTran}_i + \mathrm{tranLen}_i - \sum_{j=1}^{\mathrm{numContigsFromTran}_i} \mathrm{contigLen}_j}{\mathrm{numContigsFromTran}_i}^{-1}
\end{aligned}
$$

```
               Column 0                     Column 1                     Column 2
               --------                     --------                     --------
transcript:    012345678901234567890        01234567890123456789         0123456789012345678
contigs:          xxxxxxxyyyyyzzzzzz 000
                 xxxxxxx yyyyyzzzzzz 001
                 xxxxxxxyyyyy zzzzzz 002
                 xxxxxxxyyyyyzzzzzz  003      xxxxxxxyyyyyzzzzzz 00
                xxxxxxx  yyyyyzzzzzz 011
                xxxxxxx yyyyy zzzzzz 012
                xxxxxxx yyyyyzzzzzz  013     xxxxxxx yyyyyzzzzzz 01
                xxxxxxxyyyyy  zzzzzz 022
                xxxxxxxyyyyy zzzzzz  023     xxxxxxxyyyyy zzzzzz 02
                xxxxxxxyyyyyzzzzzz   033     xxxxxxxyyyyyzzzzzz  03       xxxxxxxyyyyyzzzzzz 0
               xxxxxxx   yyyyyzzzzzz 111
               xxxxxxx  yyyyy zzzzzz 112    xxxxxxx  yyyyyzzzzzz 11
               xxxxxxx  yyyyyzzzzzz  113
               xxxxxxx yyyyy  zzzzzz 122
               xxxxxxx yyyyy zzzzzz  123    xxxxxxx yyyyy zzzzzz 12
               xxxxxxx yyyyyzzzzzz   133    xxxxxxx yyyyyzzzzzz  13      xxxxxxx yyyyyzzzzzz 1
               xxxxxxxyyyyy   zzzzzz 222    xxxxxxxyyyyy  zzzzzz 22
               xxxxxxxyyyyy  zzzzzz  223
               xxxxxxxyyyyy zzzzzz   233    xxxxxxxyyyyy zzzzzz  23      xxxxxxxyyyyy zzzzzz 2
               xxxxxxxyyyyyzzzzzz    333    xxxxxxxyyyyyzzzzzz   33      xxxxxxxyyyyyzzzzzz  3
               [same, with order xzy]       [same, with order xzy]       [same, with order xzy]
               [same, with order yxz]       [same, with order yxz]       [same, with order yxz]
               [same, with order yzx]       [same, with order yzx]       [same, with order yzx]
               [same, with order zxy]       [same, with order zxy]       [same, with order zxy]
               [same, with order zyx]       [same, with order zyx]       [same, with order zyx]

               Column 3                     Column 4                     Column 5
               --------                     --------                     --------
transcript:    012345678901234567890        01234567890123456789         0123456789012345678
contigs:          xxxxxxxyyyyyyyyyyy 000
                 xxxxxxx yyyyyyyyyyy 001
                 xxxxxxxyyyyyyyyyyy  002      xxxxxxxyyyyyyyyyyy 00
                xxxxxxx  yyyyyyyyyyy 011
                xxxxxxx yyyyyyyyyyy  012     xxxxxxx yyyyyyyyyyy 01
                xxxxxxxyyyyyyyyyyy   022     xxxxxxxyyyyyyyyyyy  02       xxxxxxxyyyyyyyyyyy 0
               xxxxxxx   yyyyyyyyyyy 111
               xxxxxxx  yyyyyyyyyyy  112    xxxxxxx  yyyyyyyyyyy 11
               xxxxxxx yyyyyyyyyyy   122    xxxxxxx yyyyyyyyyyy  12      xxxxxxx yyyyyyyyyyy 1
               xxxxxxxyyyyyyyyyyy    222    xxxxxxxyyyyyyyyyyy   22      xxxxxxxyyyyyyyyyyy  2
               [same, with order yx]        [same, with order yx]        [same, with order yx]
```

Figure 7: Each column shows a transcript, in the first line, and all possible contig placements, in the remaining line. Each contig placement is shown both (i) according to which bases of the transcript each contig covers, with contig 1 = `xxxxxxx`, contig 2 = `yyyyy`, and contig 3 = `zzzzzz`, and also (ii) according to its vector of relative positions of uncovered bases, where 0 = “before contig 1” and *i* = “after contig *i*. For example, “ `xxxxxxxyyyyy zzzzzz`” and “`002`” represent the same contig placement. We get the following number of contig placements:

- In column 0, where $n = 4, k = 3$, there are $20 = \binom{4+3-1}{3}$ placements.
- In column 1, where $n = 4, k = 2$, there are $10 = \binom{4+2-1}{2}$ placements.
- In column 2, where $n = 4, k = 1$, there are $4 = \binom{4+1-1}{1}$ placements.
- In column 3, where $n = 3, k = 3$, there are $10 = \binom{3+3-1}{3}$ placements.
- In column 4, where $n = 3, k = 2$, there are $6 = \binom{3+2-1}{2}$ placements.
- In column 5, where $n = 3, k = 1$, there are $3 = \binom{3+1-1}{1}$ placements.

4.4.6 Summary of the distribution of $(\mathrm{Contigs}, \mathrm{Coverage}, \mathrm{ContigsFromTrans})|\{\mathrm{Reads}, \mathrm{Trans}, \mathrm{Expr}\}$

In summary,

$$\begin{aligned}
&P(\mathrm{contigs}, \mathrm{coverage}, \mathrm{contigsFromTrans}|\mathrm{reads}, \mathrm{trans}, \mathrm{expr})\\
&= P(\mathrm{contigs}, \mathrm{coverage}, \mathrm{contigsFromTrans}|\mathrm{trans}, \mathrm{expr})\\
&= P(\mathrm{numContigsFromTran}|\mathrm{trans}, \mathrm{expr})\\
&\quad \cdot P(\mathrm{contigsFromTran}|\mathrm{numContigsFromTran}, \mathrm{trans}, \mathrm{expr})\\
&\quad \cdot P(\mathrm{coverage}|\mathrm{contigsFromTran}, \mathrm{numContigsFromTran}, \mathrm{trans}, \mathrm{expr})\\
&\quad \cdot P(\mathrm{contigLen}|\mathrm{coverage}, \mathrm{contigsFromTrans}, \mathrm{trans}, \mathrm{expr})\\
&\quad \cdot P(\mathrm{contig}|\mathrm{contigLen}, \mathrm{coverage}, \mathrm{contigsFromTrans}, \mathrm{trans}, \mathrm{expr})\\
&= \left[\prod_{i=1}^{\mathrm{numTrans}} \mathrm{Binomial}(\mathrm{numContigsFromTran}_i|\mathrm{tranLen}_i, \alpha_i(1-\alpha_i)^{\mathrm{readLen}-\mathrm{ws}})\right]\\
&\quad \cdot \left[\binom{\mathrm{numContigs}}{\mathrm{numContigsFromTran}_i : i = 1, \ldots, \mathrm{numTrans}}^{-1}\right]\\
&\quad \cdot \left[\prod_{i=1}^{\mathrm{numTrans}} \frac{\prod_{j\in\mathrm{contigsFromTran}_i} \mathrm{Geometric}(\mathrm{coverage}_j|\gamma_i)}{\mathrm{NegativeBinomial}(\mathrm{expr}_i|\mathrm{numContigsFromTran}_i, 1-\gamma_i)}\right]\\
&\quad \cdot \left[\prod_{i=1}^{\mathrm{numTrans}} \frac{\prod_{j\in\mathrm{contigsFromTran}_i} \binom{\mathrm{coverage}_j-1}{\mathrm{contigLen}_j-\mathrm{readLen}}_{\mathrm{readLen}+1}}{\sum_{L=0}^{\mathrm{tranLen}_i-\mathrm{numContigsFromTran}_i\cdot\mathrm{readLen}} \binom{\mathrm{expr}_i-\mathrm{numContigsFromTran}_i}{L}_{\mathrm{readLen}+1}}\right]\\
&\quad \cdot \left[\prod_{i=1}^{\mathrm{numTrans}} \binom{\mathrm{numContigsFromTran}_i + \mathrm{tranLen}_i - \sum_{j=1}^{\mathrm{numContigsFromTran}_i} \mathrm{contigLen}_j}{\mathrm{numContigsFromTran}_i}^{-1}\right]
\end{aligned}$$