APPENDIX

A.1 Transcript Discovering Problem is NP-hard

Transcripts Discovering (TD) Problem: Given a de Bruijn graph $G(V,E)$ with a set of vertices V and edges E, a set of paired-end reads $P = \{(v_i,v_j)\}$, v_i, $v_j \in V$, an insert distance d and error s, find t paths in G with the maximum number of supporting paired-end reads $P' \subseteq P$. A path p has a supporting paired-end read (v_i,v_j) iff p contains vertices v_i and v_j and the distance between v_i and v_j in p is between $d - s$ and $d + s$.

Minimum Set Cover (MSC) Problem: Given a set of elements $U = \{1, 2, \ldots, m\}$, n set of elements $S = \{S_i\}$, $S_i \subseteq U$, and a value k, find at most k sets S' in S such that $\cup_{Si\in S'} S_i = U$.

The TD Problem can be proved to be NP-hard by reducing the MSC Problem, a NP-hard complete problem, to the TD problem.

Given an instance of the MSC problem, we can construct an instance of the TD problem. For simplicity, we assume that each element appears in at most 4 sets. We will describe the case when the cardinality of some sets in S are larger than 4 later. First, construct a de Bruijn graph with $n(m+1) + 2m$ vertices as follows. There are $m + 1$ vertices $u_{i,0}, \ldots, u_{i,m}$ representing each set S_i and two vertices $v_{j,0}$ and $v_{j,1}$ representing each element j. There is an edge from $v_{j,0}$ to $v_{j,1}$. In order to indicate whether set S_i contains element j, we either introduce a path crossing $u_{i,j-1}$, $v_{j,0}$, $v_{j,1}$ and $u_{i,j}$ (S_i contains element j) or connect $u_{i,j-1}$ and $u_{i,j}$ directly (S_i does not contain element j). Thus, if set S_i contains element j, there is an edge from $u_{i,j-1}$ to $v_{j,0}$ and an edge from $v_{j,1}$ to $u_{i,j}$. Otherwise, there is an edge from $u_{i,j-1}$ to $u_{i,j}$.

For each pair of vertices $v_{j,0}$ and $v_{j,1}$, there is a paired-end reads $(v_{j,0}, v_{j,1})$. For each pair of vertices $u_{i,j-1}$ and $u_{i,j}$, there are m^2 paired-end read $(u_{i,j-1}, u_{i,j})$. We can solve the MSC problem by solving the TD problem with maximum number of paths $t = k$, insert distance $d = 2$ and error $s = 1$. Since the number of paired-end reads $(u_{i,j-1}, u_{i,j})$ is much larger than the number of paired-end reads $(v_{j,0}, v_{j,1})$, each of the k selected path should pass through one set of points $u_{i,0}, \ldots, u_{i,m}$ representing a set S_i, i.e. with km^3 paired-end reads support. The number of elements in the union of the corresponding selected S_i can be determined from the number of supports from $(v_{j,0}, v_{j,1})$. If there are at most k paths with $km^3 + m$ paired-end reads support, the MSC problem can be solved using the set corresponding to the starting vertices of the k paths. Otherwise, the MSC problem cannot be solved.

Theorem 1: If the MSC problem can be solved, there are k paths with $km^3 + m$ paired-end reads support.

Proof: Let S_i be a selected set in the solution of the MSC problem containing a set of elements D_i, there is a simple path passing through $u_{i,0}, \ldots, u_{i,m}, v_{j,0}, v_{j,1}$ for $j \in D_i$ with $m^3 + |D_i|$ paired-end reads support. Since the selected k sets cover all elements, the corresponding k paths will have $km^3 + |\cup D_i| = km^3 + m$ paired-end reads support. □

Theorem 2: If the TD problem can be solved with $km^3 + m$ paired-end reads support, there are k sets covering all elements in the MSC problem.

Proof: Let P_i be a selected path in the TD problem, since the de Bruijn graph is a directed acyclic graph, P_i can pass though at most $m + 1$ vertices $u_{x,y}$ and $2mv_{p,q}$. Assume P_i pass through some vertices $u_{x,y}$ and $u_{x',y}$ with $x \neq x'$, it can get at most $m^2(m - 1) + m = m^3 - m^2 + m$ paired-end reads support, otherwise, it can get at most $m^3 + m$ paired-end reads support. Since $(k - 1)(m^3 + m) + (m^3 - m^2 + m) = km^3 + km - m^2 < km^3 + m$, each selected path P_i should go through $m + 1$ vertices $u_{i,0}, \ldots, u_{i,m}$ representing the same set in the MSC problem. As there are $km^3 + m$ paired-end reads support in total, for each vertex $v_{j,0}$, there is at least one selected path $v_{j,0}$ (i.e. the corresponding selected set contains element j). There are k sets covering all elements in the MSC problem. □

There are at most 4 outgoing and incoming edges for each vertex, representing the 4 possible nucleotides, in the de Bruijn graph. When some elements occur in more than 4 sets, say element j occurs in more than 4 sets S_i, we cannot construct an edge from $u_{i,j-1}$ to $v_{j,0}$ and an edge from $v_{j,1}$ to $u_{i,j}$ directly. Instead, we can construct edges from $u_{i,j-1}$ to some dummy vertices and edges from these dummy vertices to $v_{j,0}$ for reducing the in-degree of $v_{j,0}$. Similarly, we can construct edges from $v_{j,1}$ to some dummy vertices and edges from these dummy vertices to $u_{i,j}$ for reducing out-degree of $v_{j,0}$. The value of insert distance d and error s can be increased accordingly.

Table 8. Statistics of assembly result of Oases-M for simulated data and real data (completeness = 0.8)

	contigs #	avg. len. (nt)	total len. (nt)	reconstructed transcripts #	correct contigs #	sens.	spec.
Simulated data	70006	1931	109M	16991	32913	75.85%	47.01%
Real data	94248	1092	102944889	8796	17396	39.26%	18.46%

Table 9. Expression level distribution of reconstructed transcripts of Oases-M for simulated data and real data (completeness = 0.8)

Depth	(0,5)	[5,10)	[10, 15)	[15,20)	≧ 20
Total number of transcripts	5943	5011	2943	1857	6646
Simulated data	1648	3224	2461	1606	5481
Real data	577	1221	1251	965	4781