frontiers in Genetics

ORIGINAL RESEARCH
published: 29 October 2019
doi: 10.3389/fgene.2019.01020

A New Algorithm for Identifying Genome Rearrangements in the Mammalian Evolution

Juan Wang[1], *Bo Cui*[1], *Yulan Zhao*[1] *and Maozu Guo*[2,3]*

[1] School of Computer Science, Inner Mongolia University, Hohhot, China, [2] School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China, [3] Beijing University of Civil Engineering and Architecture, Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing, China

Genome rearrangements are the evolutionary events on level of genomes. It is a global view on evolution research of species to analyze the genome rearrangements. We introduce a new method called RGRPT (recovering the genome rearrangements based on phylogenetic tree) used to identify the genome rearrangements. We test the RGRPT using simulated data. The results of experiments show that RGRPT have high sensitivity and specificity compared with other tools when to predict rearrangement events. We use RGRPT to predict the rearrangement events of six mammalian genomes (human, chimpanzee, rhesus macaque, mouse, rat, and dog). RGRPT has recognized a total of 1,157 rearrangement events for them at 10 kb resolution, including 858 reversals, 16 translocations, 249 transpositions, and 34 fusions/fissions. And RGRPT has recognized 475 rearrangement events for them at 50 kb resolution, including 332 reversals, 13 translocations, 94 transpositions, and 36 fusions/fissions. The code source of RGRPT is available from https://github.com/wangjuanimu/data-of-genome-rearrangement.

Keywords: genome rearrangements, mammal, phylogenetic tree, evolution, algorithm

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Yungang Xu,
University of Texas Health Science Center at Houston, United States
Zhen Tian,
Zhengzhou University, China
Wei Lan,
Guangxi University, China

***Correspondence:**
Maozu Guo
guomaozu@bucea.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *02 July 2019*
Accepted: *24 September 2019*
Published: *29 October 2019*

Citation:
Wang J, Cui B, Zhao Y and Guo M (2019) A New Algorithm for Identifying Genome Rearrangements in the Mammalian Evolution. Front. Genet. 10:1020. doi: 10.3389/fgene.2019.01020

INTRODUCTION

The rapid development of sequencing technologies makes the phylogenetic analysis from the level of whole genome possible. A studied genome is represented as a line of conserved segments (called syntenic blocks). The genome rearrangements of species are changes of syntenic block orderings and losing of sequence blocks. These events include reversal, translocation, transposition, fusion, fission, and so on (Xu et al., 2017; Cheng et al., 2019; Dong et al., 2018). The research on genome rearrangements is mainly three aspects.

One is the computation of evolutionary distance between two species by considering genome rearrangements. Researchers have proposed a lot of metric for measuring the dissimilarity of evolution between species and a large amount of algorithms for computing the metrics. The breakpoint distance is the minimum rearrangement operations transforming one genome to the other genome, which is computed by means of breakpoint graph (Blanchette et al., 1997; Sankoff and Blanchette, 1998). There are lots of algorithms for computing breakpoint distance. In 1995, Hannenhalli and Pevzner put forward an algorithm with $O(n^5)$ time complexity to compute the breakpoint distance just considering reversal events (Hannenhalli and Pevzner, 1999). Later, Kaplan improved the algorithm to time complexity $O(n^5)$ (Kaplan et al., 2000). In 1996, Hannenhalli designed an algorithm with $O(n^3)$ time complexity to compute it by

considering translocation events (Hannenhalli, 1995). In 2001, Zhu et al. improved the algorithm to time complexity $O(n^2 logn)$ (Zhu and Ma, 2002). And then Zhu et al. devised an algorithm with $O(n^2)$ time complexity (Liu et al., 2004). The DCJ distance is introduced by Yancopoulos et al. (Sophia et al., 2005), which uses the double cut and join (DCJ for short) operation to model rearrangement events, such as reversal, translocation, transposition, fusion, and fission in an unified way. Yancopoulos et al. first propose a method to compute the DCJ distance by considering only translocations and reversals on linear chromosomes (Sophia et al., 2005). Paper (Lu et al., 2006) has proposed an $O(n^2)$ time algorithm to compute the distance by considering the fusions and fissions between circular unsigned chromosomes. Unimog (Hilker et al., 2012) is software for computing DCJ distance which implements lots of algorithms (Erdös et al., 2011; Jakub et al., 2011). SoRT is a tool to compute breakpoint distance and the DCJ distance for linear/circular multi-chromosomal gene orders (Yen-Lin et al., 2010). SCJ distance (Feijão and Meidanis, 2011) is defined using the single cut and join (SCJ for short) operations, which is in analogy to DCJ measure. The distance can be computed by a speedily computable.

Two is the reconstruction of the ancestral gene orders by using the genomes of extant species. Ma et al. (Ma et al., 2006) use maximum parsimony principle to recover reliably ancestral genomes starting from phylogenetic tree and adjacent genes in genome and make the probabilistic reconstruction accuracy analysis for the six mammalian genome (human, mouse, rat, dog, opossum, and chicken) based on the improved Jukes–Cantor model. PMAG utilized the Bayesian theorem in the probabilistic framework to infer ancestral genomes (Yang et al., 2014). Multiple Genome Rearrangements (MGR) recovers the ancestral genome by minimizing the rearrangement distance (Bourque and Pevzner, 2002). Multiple Genome Rearrangements and Ancestors (MGRA) is developed to reconstruct ancestral genomes based on multiple breakpoint graphs and is used to analyze rearrangement evolutionary events of seven mammalian genomes (human, chimpanzee, macaque, mouse, rat, dog, and opossum) (Alekseyev and Pevzner, 2009). Decostar (Duchemin et al., 2017) is a software which reconstructs neighborhood relations of ancestral genes aiming at reconstructing the organization of ancestral genomes.

Three is the recognition of the rearrangement events of existing species. Efficient Method to Recover Ancestral Events (EMRAE) is an algorithm which can recognize rearrangement events in evolution described by phylogenetic tree by means of adjacent genes in genomes (Zhao and Bourque, 2009).

MATERIALS AND METHODS

Preliminaries

A genome is composed of several chromosomes, and each chromosome is an ordering of syntenic blocks. For convenience, each syntenic block is recorded by an integer, so a chromosome is represented by a signed permutation $X=c_1c_2\cdots g_n$, where $c_i(1\leq i\leq n)$ is an integer representing a syntenic block, its sign is assigned with the orientation that is either positive (recorded by c_i) or negative (recorded by $-c_i$). The chromosome $X=c_1c_2\cdots c_n$ is the same as $-X = -c_n - c_{n-1}\cdots - c_1$.

A reversal $r(i, j)$ $(i \leq j)$ converts chromosome $X=c_1c_2\cdots c_n$ into a new chromosome $X'=c_1c_2\cdots-c_j-c_{j-1}\cdots-c_{i+1}-c_ic_{j+1}\cdots c_n$, where the reversal is from c_i to c_j.

A translocation event breaks two chromosomes into four segments and then reconnects them into two new chromosomes. Given two chromosomes $X = X_1X_2$ and $Y = Y_1Y_2$, where $X_1=x_1x_2\cdots x_{i-1}$,$X_2=x_ix_{i+1}\cdots x_m$,$Y_1=y_1y_2\cdots y_{j-1}$, and $Y_2=y_jy_{j+1}\cdots y_n$, a translocation is represented by $tl(i,j)$. X_1 and Y_1 are exchanged to form two new chromosomes $X'=Y_1X_2$ and $Y'=X_1Y_2$, or X_1 and Y_2 are exchanged to form two new chromosomes $X" = -Y_2X_2$ and $Y" = X_1 - Y_1$.

A transposition event is to exchange two adjacent fragments on one chromosome into a new chromosome. A transposition is represented by $tp(i, j, k)$, i.e., the fragment $c_i\cdots c_j$ of one chromosome inserted into after c_k. If c_k is on the same chromosome ($k > j$ or $k < i$), then the transposition $tp(i, j, k)$ is called intra-chromosomal; otherwise, it is inter-chromosomal. Given a chromosome $X=c_1c_2\cdots c_ic_{i+1}\cdots c_{j-1}c_j\cdots c_k\cdots c_n$ and an intra-chromosomal transposition, X is converted into $X'=c_1c_2\cdots c_kc_ic_{i+1}\cdots c_jc_{k+1}\cdots c_n$.

A fusion event is to connect two chromosomes into a new chromosome. The fusion acting on chromosomes X_1 and X_2 is represented by $fu(X_1, X_2)$ and forming a new chromosome X_1X_2 or X_1-X_2. A fission is to split a chromosome into two new chromosomes. A fission acting on the chromosome $X = X_1X_2$ is represented by $fi(X)$ and forming two new chromosomes X_1 and X_2 (where X_1 and X_2 are non-empty segments).

An adjacency $a(c_i,c_{i+1})$ of genome X is two adjacent integers in one chromosome of X. $a(c_i,c_{i+1})$ is the same as $a(-c_{i+1},-c_i)$. For example, all adjacencies on chromosome $X = 1,234$ are $a(1, 2)$, $a(2, 3)$, and $a(3, 4)$. For a set of genomes S, an adjacency a is effective w.r.t. S if it belongs to at least one genome and not all genomes. For example, two uni-chromosomal genomes G_1 and G_2, the chromosome $X = 1,234$ of G_1 and the chromosome $Y = 1 - 3 - 24$ of G_2, then all effective adjacencies w.r.t. G_1 and G_2 are $a(1, 2)$, $a(2, 3)$, $a(3, 4)$, $a(1, -3)$, and $a(-2, 4)$.

EMRAE

Given a phylogenetic tree T describing the evolution of the genomes G, EMRAE first computes all effective adjacencies w.r.t. G. Then, it predicts the rearrangement events for each edge of T by means of inference rules (will be introduced in the following).

Figure 1 shows a reversal $r(2, 3)$ during the evolution from A to B, where A and B are two uni-chromosomal genomes, and the chromosomes are $X = 1,234$ and $Y = 1 - 3 - 24$, respectively. The set of genomes will be divided into two subsets recorded by S_A and S_B after removing the edge e from T. Suppose there is not any rearrangement events inside S_A and S_B. Then, adjacencies $a(1, 2)$ and $a(3, 4)$ can be found in each genome of S_A and not in any one genome of S_B; $a(1,-3)$ and $a(-2,4)$ can be

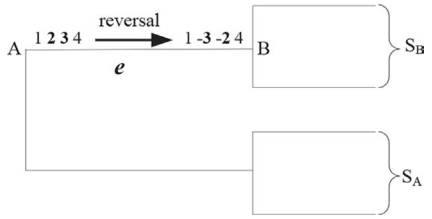

FIGURE 1 | A reversal ***r*** (2, 3) during the evolution from *A* to *B*; *S*\s\do5 **(A)** and *S*\s\do5 **(B)** are two subsets of all leaves species divided by the edge *e*.

found in each genome of S_B and not in any one genome of S_A. In turn, we can utilize the four adjacencies $a(1, 2)$, $a(3, 4)$, $a(1, -3)$, and $a(-2,4)$ to identify a reversal $r(2, 3)$ occurring on the edge *e*. The EMRAE method infers the rearrangement events by means of the similar rules.

Let $e = (A, B)$ be an edge of T, $G=\{G_1,G_2,\cdots,G_m\}$the genomes of leaves, and $a_1,a_2,\cdots a_i$ the children of A and $b_1,b_2,\cdots b_j$ the children of B. EMRAE first selects a number of adjacencies as candidate adjacencies $Ca(e,A)$ for edge e and node A according the following steps.

1. Find the adjacencies are in each genome of S_A and not in any one genome of S_B, then put them to $Ca(e, A)$;
2. If A is an internal node, find all edges connected with A except e and record them with $e_1,e_2,\cdots,e_k$. For each $e_i=(u_i,A)(1\le i\le k)$, G can be divided into two parts after removing e_i, S_{ui} is the part not including A.
   a. Find the adjacencies that are in one genome of each S_{ui} $(1 \le i \le k)$ and not in any one genome of S_B, then put them to $Ca(e,A)$;
   b. Compute $Ca(e_i, u_i)$ and $Ca(e_i,u)(1\le i\le k)$. For each one $Ca(e_i, u_i)$, find the adjacency a_1 from $Ca(e_i, u_i)$, such that a_1 is not overlap gene with any one adjacency in $Ca(e_i, u)$, a_1 has overlap gene with one adjacency a_2 in each $Ca(e_j,u_j)(1\le j\ne i\le k)$, and a_2 has overlap gene with at least one adjacency in $Ca(e_j, u)$, then put *a*\s\do5(1) to $Ca(e, u)$.

EMRAE then infers rearrangement from $Ca(e, A)$ and $Ca(e, B)$ for edge $e = (A, B)$ with the help of inference rules in the following section. From the definitions of genome rearrangements, we find that each genome rearrangement can change several adjacencies. For example, each reversal $r(i, j)(i \le j)$ can change two adjacencies $a_1=a(c_{i-1},c_i)$ and $a_2=a(c_j,c_{j+1})$ into $b_1 = a(c_i-1, -c_j)$ and $b_2=a(-c_i,c_{j+1})$. Based on those facts, we obtain the inference rules introduced in the following section.

Inference Rule

Let $e = (A,B)$ be an edge of the phylogenetic tree T. Given adjacencies $a_1 = a(c_1-1, c_i)$, $a_2 = a(c_j, c_{j+1})$ in $Ca(e,A)$ and $b_1=a(c_{i-1},-c_j)$, $b_2=a(-c_i,c_{j+1})$ in $Ca(e,B)$, EMRAE infers a reversal $r(i,j)$ from A to B if all genomes are uni-chromosomal or a_1, a_2 are in the same chromosome in S_A and b_1, and b_2 are in the same chromosome in S_B. Otherwise, we infer a translocation $tl(i, j)$. Similarly, given adjacencies $a_1=a(c_{i-1},c_i)$, $a_2=a(c_jc_{j+1})$ in $Ca(e,A)$ and $b_1=a(c_{i+1},c_{j+1})$, $b_2=a(c_j,c_i)$ in $Ca(e,B)$, EMRAE infers a translocation $tl(i,j)$, or a reversal for a_1, a_2 in $Ca(e,A)$ and adjacencies b_1, b_2 in $Ca(e,B)$.

Assume that there are adjacencies $a_1=a(c_{i-1},c_i)$, $a_2=a(c_j,c_{j+1})$, and $a_3=a(c_k,c_{k+1})$ in $Ca(e,A)$ and $b_1=a(c_{i-1},c_{j+1})$, $b_2=a(c_k,c_i)$, and $b_3=a(c_j,c_{k+1})$ in $Ca(e,B)$. EMRAE can predict a transposition $tp(i,j,k)$ during the evolution from A to B if all genomes are uni-chromosomal. Otherwise, suppose m genomes in S_A have a_1 and a_2, then EMRAE can predict a transposition $tp(i,j,k)$ if there are at least $m/2$ genomes such that the four integers of a_1 and a_2 on the same chromosome, or there are at least $m/2$ genomes such that the four integers of a_2 and a_3 on the same chromosome.

Assume that there is $a=a(c_i,c_j)$ in $Ca(e,A)$. EMRAE can predict a fission that splits the adjacency $a=a(c_i,c_j)$ if a is sign-compatible for each genome G_k in S_B. The fusion from A to B can be seen as a fission from B to A.

Recovering the Genome Rearrangements Based on Phylogenetic Tree

EMRAE can not identify the rearrangement occurring in the frontier of genomes. We take **Figure 2**, for example, where species A, B, and C are uni-chromosomal genomes A = 1,234, B = −2 − 134, and C = 1,234. A reversal r(1,2) has occurred in the evolution from A to B. EMRAE can compute the candidate adjacencies $a(-1,3)$ for $Ca(e_1,B)$ and $a(2,3)$ for $Ca(e_1,A)$. So, EMRAE can not infer the reversal r(1,2) on the edge e_1 according to the candidate adjacencies.

We improve EMRAE so that the improved method (called RGRPT) is able to infer the rearrangement events occurring in the frontier region. The inference rule of RGRPT is the same as that of EMRAE. The difference between RGRPT and EMRAE is that they have different candidate adjacencies. RGRPT puts 0 to the head and tail for each chromosome, so there will be added a lot of adjacencies for each genome. For example, considering the uni-chromosomal genomes X = 1,234 and Y = −2 −134, the two additional candidate adjacencies $a(0,1)$ and $a(0,-2)$ are added.

RGRPT adds candidate adjacencies in the step b of EMRAE. For each one $Ca(e_i,u_i)$ and an adjacency a_1 from $Ca(e_i,u_i)$, if there is an adjacency a_2 in each $Ca(e_j,u_j)(1\le j\ne i\le k)$ such that a_1 with a_2 has overlap gene, then put a_1 to $Ca(e,u)$.

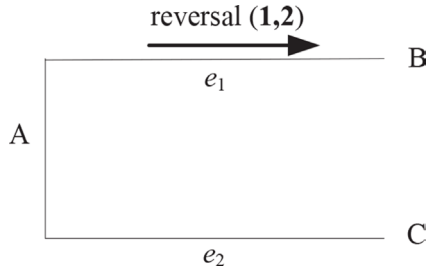

FIGURE 2 | The tree topology with two taxa (**B** and **C**).

RESULTS

All of the experiments were performed on a computer with Intel Vostro 14 2.0 GHz CPU, 4 GB RAM, and 500 GB Hard Disk Drives (HDD). The operating system was Win10 64 bit with Java 1.6 installed. RGRPT was written in Java.

We tested RGRPT with both simulated data and the practical data (i.e., real biological data) introduced by the following section.

Simulated Data

Here, we start with an uni-chromosomal genome as the ancestor, and it evolves along the phylogenetic tree with n taxa whose topology sees the **Figure 3**.

We generate two simulated data sets in order to test the affectivity of RGRPT. One of them is created from the phylogeny only with reversals events. The other data set is generated from the phylogeny with kinds of events, including reversals, translocation, transposition, fusion, and fission, and the quantity of those events is in a certain ratio. The two data sets can test the ability of methods to recover the simple and the complex evolution histories. First data set is created just using reversal events. Since the reversal on only one gene is rare (Korbel et al., 2007), we set the ratio of reversal on one gene and on more than one gene as 1:3. The number of leaves is from 3 to 10 with step 1. For each number of leaves, the ancestor genome with m gene, where m from 50 to 150 with step 10. Each edge will happen k reverse, where k is random integer number from 3 to 10. So, there are 11 groups data for each leaf number. Sensibility is the percentage of correctly predicted events in all practical events. Specificity is the percentage of correctly predicted events in all predicted events. We compute the sensibility and specificity for RGRPT and EMRAE for each group data. **Table 1** shows the average sensitivity and specificity for each leaf number. The second column of the table records the number of all events, and its last row records the average values.

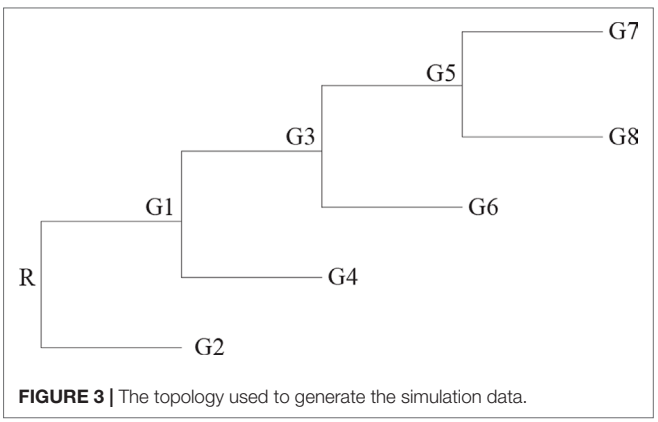

FIGURE 3 | The topology used to generate the simulation data.

Table 1 shows that RGRPT achieves higher sensibility than EMRAE, and RGRPT achieves comparable specificity with EMRAE. Obviously, RGRPT can distinguish more actually occurred events than EMRAE. So, the experimental results show that the RGRPT is more efficient than EMRAE for predicting reversal events.

Second data set is generated by using all events, i.e., reversal, translocation, transposition, fusion, and fission. The reversals are generally more than the other rearrangement events. The fusions and the fissions are very rare; so, we record the number of the two events together. Here, we set the ratio of those events as 10:2:2:0.1. The ancestor genome has 5 chromosomes and each chromosome with 100 genes. The ancestor genome evolves along the topology with four leaves (see **Figure 3**). Each edge happen k events, where k is random number from 1 to μ and μ is 6, 12, 18, and 24. For each μ, it runs 10 times; so, we can obtain 10 groups data for each μ. **Table 2** shows the average of 10 groups data for each μ. This table indicates that the RGRPT is more efficient than EMRAE for predicting all events.

Practical Data

The practical data is from the paper (Zhao and Bourque, 2009). It contains six mammalian genomes, i.e., human, chimpanzee, rhesus monkey, mouse, voles, and dog. The data are created from two different levels of resolution 10 kb and 50 kb. **Figure 4** is the tree describing the phylogeny of species. The results are shown in **Tables 3** and **4**. EM and RG represent EMRAE and RGRPT respectively, and Rev, Tloc, Tran, Fus, and Fis represent reversal,

TABLE 1 | Results of EMRAE and recovering the genome rearrangements based on phylogenetic tree algorithms in predicting reversal events.

Leaves	Reversal	Sensibility		Specificity	
		EMRAE	RGRPT	EMRAE	RGRPT
3	24	64%	76%	89%	90%
4	39	65%	76%	94%	94%
5	45	61%	72%	92%	93%
6	59	57%	66%	90%	90%
7	69	54%	65%	92%	91%
8	79	59%	80%	92%	92%
9	92	55%	63%	90%	90%
10	104	55%	62%	89%	89%
Mean		58.7%	70%	91%	91.1%

TABLE 2 | Results of EMRAE and recovering the genome rearrangements based on phylogenetic tree algorithms in predicting all events.

Events of each edge	All events	Sensibility		Specificity	
		EMRAE	RGRPT	EMRAE	RGRPT
6	19	75.8%	85.7%	95.8%	96.2%
12	29	74.2%	80.3%	97%	96.5%
18	38	53.5%	58.1%	95.4%	96.7%
24	50	47.7%	50.5%	94.9%	94.1%
Mean		62.8%	68.7%	95.8%	95.9%

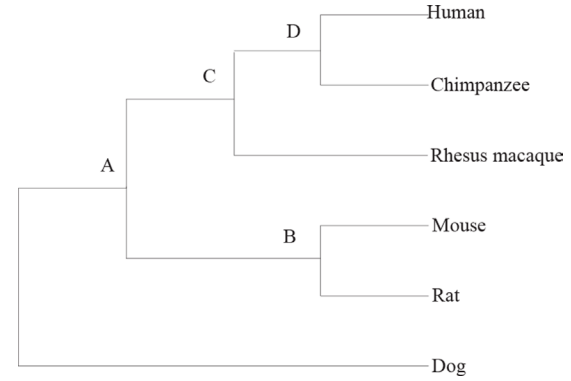

FIGURE 4 | The tree describing the phylogeny of mammalian species.

translocation, transposition, fusion, and fission, respectively. Each row in the table records the ancestor rearrangement events of the edge. For example, the values in the human row are the rearrangement events from D to human; the values in MR row are the rearrangement events from A and B.

At 10 kb resolution, the RGRPT algorithm predicts 1,157 ancestor rearrangement events, including 858 reversals, 16 translocations, 249 transpositions, and 34 fusions and fissions. It identifies 48 rearrangement events more than the EMRAE. The reversal events are in the majority in all predicted events. At 50 kb resolution, the RGRPT algorithm predicts 475 ancestor rearrangement events, including 332 reversals, 13 translocations, 94 transpositions, and 36 fusion and fissions. RGRPT identifies 21 rearrangement events more than EMRAE algorithm. The rearrangement events identified in the rat

TABLE 3 | Genome rearrangement predictions of EMRAE and recovering the genome rearrangements based on phylogenetic tree at 10 kb resolution.

Species	Rev		Tloc		Tran		Fus/Fis		Total events	
	EM	RG	EM	RG	EM	RG	EM	RG	EM	RG
Human	12	13	0	0	4	5	0	0	16	18
HC	29	32	0	0	15	15	0	1	44	48
HCP	83	84	0	0	8	10	2	8	93	102
Chimp	17	19	0	0	7	8	1	1	25	28
Rhesus	49	50	0	0	40	42	1	2	90	94
Mouse	90	95	3	3	10	13	5	5	108	116
Rat	227	233	0	0	127	129	3	3	357	365
MR	140	143	2	3	9	10	0	0	151	156
Dog	184	189	10	10	17	17	14	14	225	230
Total	831	858	15	16	237	249	26	34	1,109	1,157

TABLE 4 | Genome rearrangement predictions of EMRAE and recovering the genome rearrangements based on phylogenetic tree at 50 kb resolution.

Species	Rev		Tloc		Tran		Fus/Fis		Total events	
	EM	RG	EM	RG	EM	RG	EM	RG	EM	RG
Human	2	2	0	0	1	1	0	0	3	3
HC	19	19	0	0	4	4	1	1	24	24
HCP	27	29	0	0	5	6	2	6	34	41
Chimp	17	19	0	0	7	8	1	1	25	28
Rhesus	22	23	0	0	6	7	1	3	29	33
Mouse	25	27	3	3	0	0	5	6	33	36
Rat	128	131	0	0	65	65	5	5	198	201
MR	41	42	2	2	2	2	0	0	45	46
Dog	46	47	7	8	8	8	13	14	74	77
Total	322	332	12	13	92	94	28	36	454	475

edge are mostly in all edges either at 10 kb resolution or at 50 kb resolution. The syntenic blocks of genomes at 10 kb resolution are more than the syntenic blocks of genomes at 50 kb resolution. The fact reduces the recognized rearrangement events at 10 kb resolution that are more than the recognized rearrangement events at 50 kb resolution. Experiments show that RGRPT can recover more ancestor events than EMRAE.

DISCUSSION

This paper proposes a new method, RGRPT, to infer ancestor rearrangement events. RGRPT takes a phylogenetic tree describing the evolution of species and the genomes of species as input. Experiments on the simulated data and practical data show that RGRPT is more efficient than EMRAE and can recover more ancestor rearrangement events than EMRAE. RGRPT provides a method for us to research the genome rearrangement of species. We can use RGRPT to recognize the ancestral genome rearrangement for the evolution of other species in future (Tian et al., 2018).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: https://github.com/wangjuanimu/data-of-genome-rearrangement.

AUTHOR CONTRIBUTIONS

JW proposed and implemented the RGRPT method. JW and BC designed all experiments. All authors participated in the designing the algorithm and writing the paper.

FUNDING

The work was supported by the National Natural Science Foundation of China (61661040, 61661039, 61571163, 61532014, 61671189, 91735306, 61751104); the National Key Research and Development Plan Task of China (Grant No. 2016YFC0901902).

REFERENCES

Alekseyev, M. A., and Pevzner, P. A. (2009). Breakpoint graphs and ancestral genome reconstructions. *Genome Res.* 19 (5), 943–957.

Blanchette, M., Bourque, G., and Sankoff, D. (1997). Breakpoint phylogenies. *Genome Inform. Ser. Workshop Genome Inform.* 8, 25–34.

Bourque, G., and Pevzner, P. A. (2002). Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res.* 11 (1), 26–36.

Cheng, L., Yang, H., Zhao, H., Pei, X., Shi, H., Sun, J., et al. (2019). Metsigdis: a manually curated resource for the metabolic signatures of diseases. *Briefings Bioinf.* doi: 10.1093/bib/bbx103

Dong, S., Zhao, C., Fei, C., Liu, Y., Zhang, S., Hong, W., et al. (2018). The complete mitochondrial genome of the early flowering plant nymphaea colorata is highly repetitive with low recombination. *Bmc Genomics* 19 (1), 614–626.

Duchemin, W., Anselmetti, Y., Patterson, M., Ponty, Y., Brard, S., Chauve, C., et al. (2017). Decostar: Reconstructing the ancestral organization of genes or genomes using reconciled phylogenies. *Genome Biol. Evol.* 9 (5), 1312–1319.

Erdös, P. L., Soukup, L., and Stoye, J. (2011). Balanced vertices in trees and a simpler algorithm to compute the genomic distance. *Appl. Math. Lett.* 24 (1), 82–86.

Feijão, P., and Meidanis, J. (2011). Scj:a breakpoint-like distance that simplifies several rearrangement problems. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 8 (5), 1318–1329.

Hannenhalli, S. (1995). Polynomial-time algorithm for computing translocation distance between genomes. *Discrete Appl. Math.* 71 (1–3), 137–151.

Hannenhalli, S., and Pevzner, P. A. (1999). Transforming cabbage into turnip:polynomial algorithm for sorting signed permutations by reversals. *J. Acm* 46 (1), 1–27.

Hilker, R., Sickinger, C., Pedersen, C. N., and Stoye, J. (2012). Unimog–a unifying framework for genomic distance calculation and sorting based on dcj. *Bioinformatics* 28 (19), 2509.

Jakub, K., Robert, W., Braga, M. D. V., and Jens, S. (2011). Restricted dcj model: rearrangement problems with chromosome reincorporation. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* 18 (9), 1231–1241.

Kaplan, H., Shamir, R., and Tarjan, R. E. (2000). Faster and simpler algorithm for sorting signed permutations by reversals. *SIAM J. Comput.* 29 (3), 880–892.

Korbel, J. O., Urban A. E., Affourtit J. P., Godwin B., Grubert F., Simons J. F. et al. (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318 (5849), 420–426.

Liu, X., Zhu, D., Ma, S., Li, Z., and Wang, L. (2004). An o(n2) algorithm for sorting oriented genomes by translocations. *Chin. J. Comput.* 27 (10), 1354–1360.

Lu, C. L., Huang, Y. L., Wang, T. C., and Chiu, H. T. (2006). Analysis of circular genome rearrangement by fusions, fissions and block-interchanges. *Bmc Bioinf.* 7 (1), 295.

Ma, J., Zhang, L., Suh, B., and e. a. Raney, B. (2006). Reconstructing contiguous regions of an ancestral genome. *Genome Res.* 16 (12), 1557–1565.

Sankoff, D., and Blanchette, M. (1998). Multiple genome rearrangement and breakpoint phylogeny. *J. Comput. Biol.* 5, 555–570.

Sophia, Y., Oliver, A., and Richard, F. (2005). Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics* 21 (16), 3340–3346.

Tian, Z., Teng, Z., Cheng, S., and Guo, M. (2018). Computational drug repositioning using meta-path-based semantic network analysis. *BMC Syst. Biol.* 12 (S9), 134.

Xu, Y., Wang, Y., Luo, J., Zhao, W., and Zhou, X. (2017) Deep learning of the splicing (epi)genetic code reveals a novel candidate mechanism linking histone modifications to esc fate decision. *Nucleic Acids Res.* 45 (21), 12100–12112.

Yang, N., Hu, F., Zhou, L., and Tang, J. (2014). Reconstruction of ancestral gene orders using probabilistic and gene encoding approaches. *PLoS One* 9 (10), e108796.

Yen-Lin, H., Chen-Cheng, H., Chuan Yi, T., and Chin Lung, L. (2010). Sort2: a tool for sorting genomes and reconstructing phylogenetic trees by reversals, generalized transpositions and translocations. *Nucleic Acids Res.* 38 (Web Server issue), W221–W227.

Zhao, H., and Bourque, G. (2009). Recovering genome rearrangements in the mammalian phylogeny. *Genome Res.* 19 (5), 934–942.

Zhu, D., and Ma, S. (2002). An improved algorithm for the translocation sorting problem of genomes. *Chin. J. Comput.* 25 (2), 189–196.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.