Supplementary Material

RECOVERING GENOME REARRANGEMENTS IN THE MAMMALIAN PHYLOGENY

Hao Zhao and Guillaume Bourque[†]

[†] bourque@gis.a-star.edu.sg

Supplementary Text

EMRAE Pseudo-code

Algorithm EMRAE (G_1,G_2, ..G_m, T)

Input: Genomes G_1,G_2,..,G_m, and their phylogenetic tree T

Output: Inferred events on every edge e in T

1. **for** each edge e = (A,B) in T **do**
2. Compute conserved adjacencies CA(e,A) and CA(e,B)
3. **for** each edge e = (A,B) in T **do**
4. Refine CA(e,A) and CA(e,B)
6. **for** each edge e = (A;B) in T **do**
7. Infer every possible reversal *r* and translocation *tloc* and remove the 4 related adjacencies from CA(e,A) and CA(e,B)
8. **for** each edge e = (A,B) in T **do**
9. Infer every possible transposition *t* and remove the 6 related adjacencies from CA(e,A) and CA(e,B)
10. Infer every possible fusion and fission

Localization of predicted rearrangement events

The events predicted by EMRAE are partial in the sense that a predicted event is only represented by the adjacencies associated with it and these adjacencies do not have an orientation. We will now show how to define and check if a predicted event can be localized on a particular genome, we such call such events *actual*. Assume a predicted

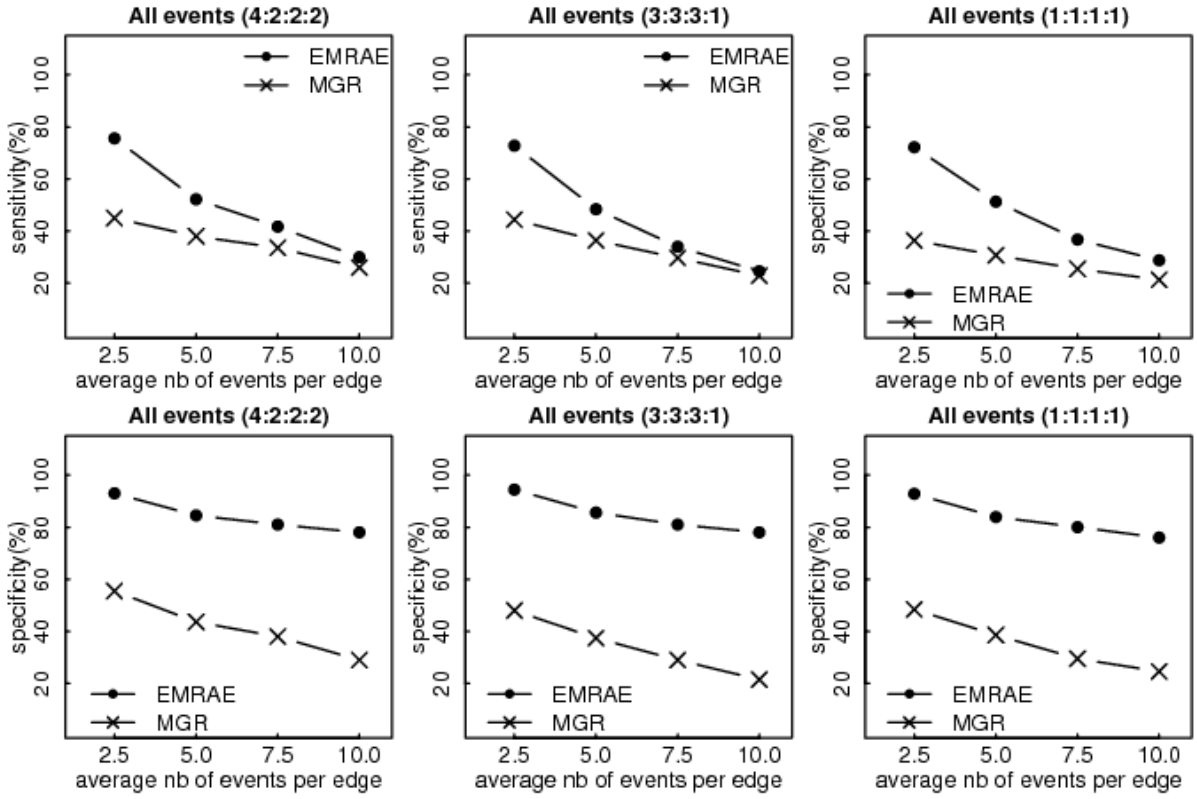
reversal from the ancestor A to B on an edge e = (A, B) affects four blocks 1, 2, 3 and 4 and is represented by a(1, -3), a(-2, 4) at the side A and their counterparts a(1, 2), a(3, 4) at the leaf B. Note that though conserved adjacencies have no orientations, if B is a leaf genome, then the orders and orientations of the four blocks 1, 2, 3, and 4 on B are unambiguous. We perform the reversal r on a(1, 2) and a(3, 4) from B back to its ancestor A. If the resulted genome includes their counterparts a(1, -3) and a(-2, 4), then we call the reversal r is an *actual reversal* from the ancestor A to its leave B on the edge e = (A, B) . Although it is natural that a predicted reversal can be an actual one, this does not always happen. In the example above, again we assume the reversal r is associated with the same four adjacencies. Further we assume that the chromosome affected by r of genome B is …1 2 … -4 -3…. Then genome B contains the two adjacencies a(1, 2) and a(3, 4). Since adjacencies have no orientations, a(-4, -3) is equal to a(3, 4). But this time when we perform the reversal from B to A, then the two adjacencies a(1, 2) and a(3, 4) are transformed into a(1, 4) and a(-2, -3) at A, different from the expected counterparts a(1, -3) and a(-2, 4). We call such a reversal *r* as an ambiguous prediction. We can check predicted transpositions similarly. Translocations can be always directly performed since it can exchange genomic content in two ways (see the definition of translocations). The simpler fusion/fission is only associated with one conserved adjacency and can always be actual ones (see our Inference Rules). In our analysis of the predictions, we remove ambiguous predictions and only focus on actual events. For predictions on internal edges, our definitions are similar but a little more stringent. In the above example, if A is the ancestor of B, and B is the common ancestor of human, chimp, then we call a reversal

r on e = (A, B) only if it can be performed directly from both human and chimp to B. See Sup. Table 4.

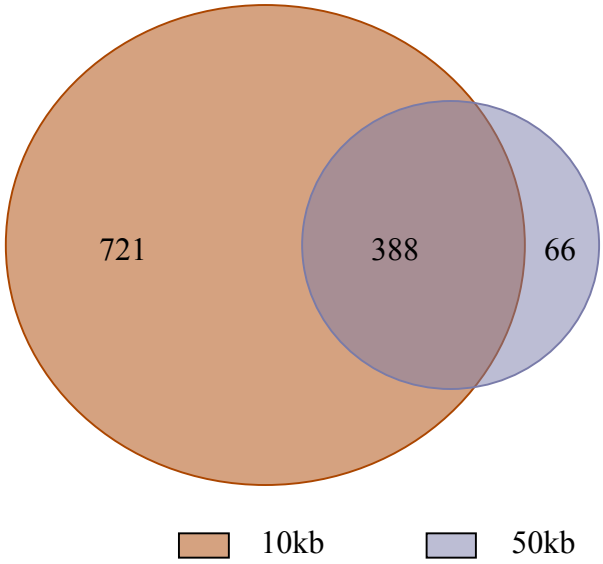
Evaluation of MGR's ability to predict transpositions

MGR does not model transpositions. As mentioned in the main text, to evaluate MGR's prediction of transpositions, we used three MGR's consecutive reversals to mimic an intra-chromosomal transposition (Zhao and Bourque 2007). Based on a similar idea, we used two translocations to mimic an inter-chromosomal transposition, with which a consecutive segment of blocks are moved to a different chromosome. For example, assume an inter-transposition *t* acting on two chromosomes chr1 = 1 2 3 4 5 6 and chr2 = 7 8 9 10 picks up the segment 3 4 from chr1 and places it after the block 8 in chr2. Then *t* leads to two new chromosomes chr1' = 1 2 5 6 and chr2' = 7 8 3 4 9 10. The following is a possible way to mimic *t* with two translocations. In the first step, a translocation $tloc_1$ is performed to exchange the segment 3 4 5 6 of chr1 with 9 10 of chr2 and it leads to two intermediate chromosomes chr1'' = 1 2 9 10 and chr2'' = 7 8 3 4 5 6. In the second step another translocation $tloc_2$ exchanges 9 10 of chr1'' with 5 6 of chr2''. Thus the two consecutive translocations $tloc_1$ and $tloc_2$ work equivalently as the transposition *t* and get the target chromosomes chr1' = 1 2 5 6 and chr2' = 7 8 3 4 9 10.

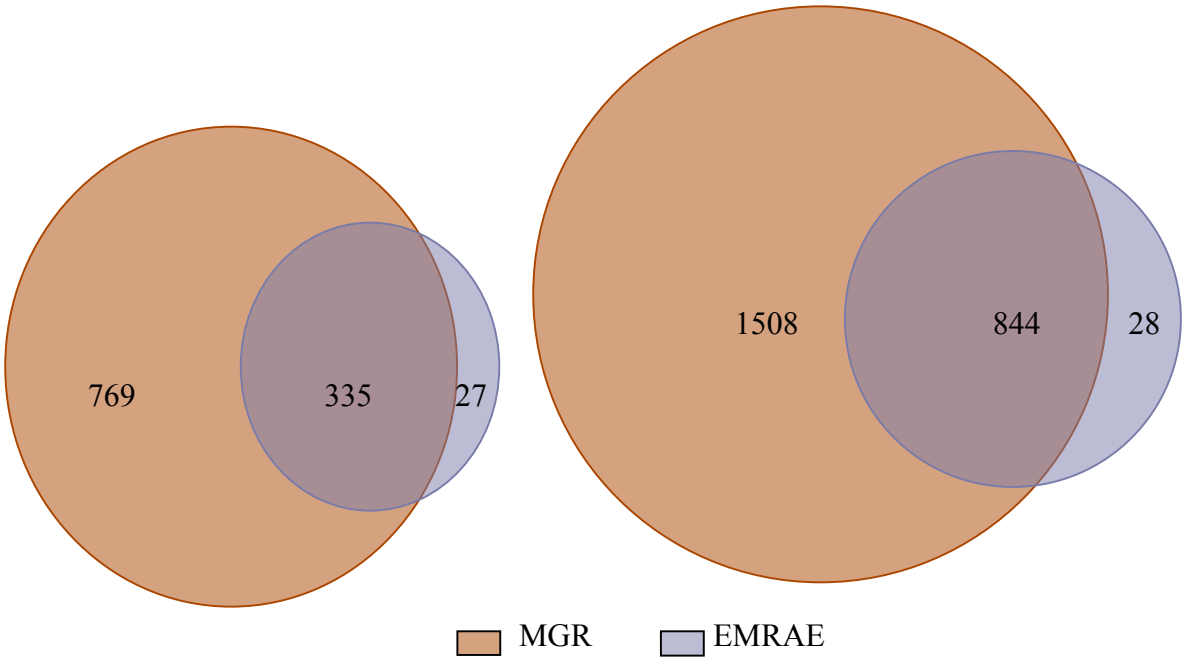
Supplementary Figures



Supplementary Figure 1. Comparison of EMRAE and MGR's predictions for the All events model. The results are shown for 3 ratios of reversals, translocations, transpositions, fusion/fissions: 4:2:2:2, 3:3:3:1 and 1:1:1:1.



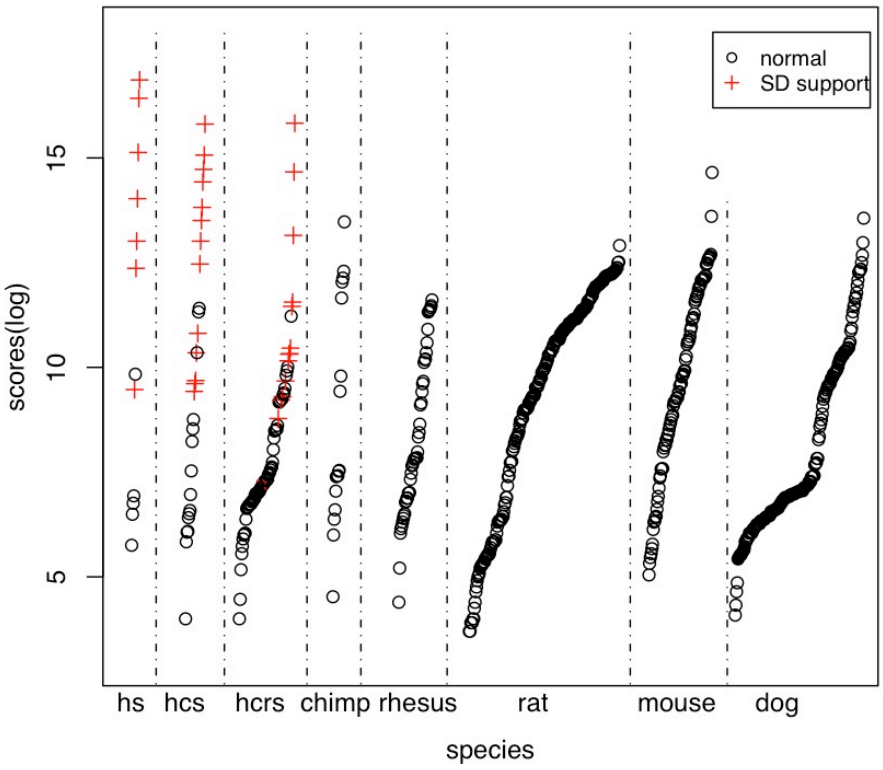
Supplementary Figure 2. Comparison of EMRAE's predictions between the 10Kb and 50Kb datasets.



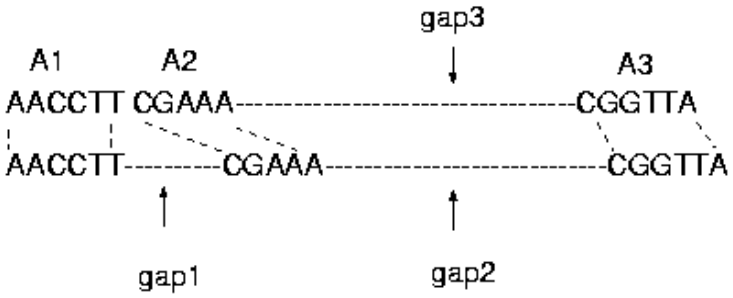
(A) Comparison on the 50-kb dataset

(B) Comparison on the 10-kb dataset

Supplementary Figure 3. Comparison between EMRAE and MGR excluding transpositions.



Supplementary Figure 4. Best Blast scores between the pairs of breakpoint regions associated with rearrangement events in the different lineages. The scores are plotted in a log-scale. Those primate reversals supported with pairs of SDs are labeled with "+".



Supplementary Figure 5. An example of splitting chained alignments. In this example, the chain consists of three un-gapped alignments A1, A2 and A3. And gap1 is smaller than 300bps, while gap2 or gap3 is larger than 300 bps. Then we will break the chain into two independent alignments: A1 and A2 are merged into a single alignment since gap1 is

ignored; gap2 and gap3 are removed and thus A3 is separated into an independent alignment.

Supplementary Tables

Lineage	Bp1	Genes with bp1	Bp2	Genes with bp2
human	chr5:68925902-69361045	uc003jxm.2,uc003jxp.2*,uc010ixq.1, uc010ixr.1,uc003jxr.1,uc003jzg.2, uc003jxx.2,uc003jxz.2*,uc003jya.1*, uc003jyb.1*,	chr5:69457387-70782081	
human	chr11:131981281-132001503	uc009zcy.1,uc001qgs.1,uc001qgt.1, uc001qgu.1,	chr11:132003494-132023494	
human	chr13:113373993-113636378	uc001vuh.1,	chr13:113651104-113758051	
human	chrX:101038511-101425486	uc004eit.1,uc004eiv.2,uc004eiu.2, uc004eiw.2,uc004eix.2,	chrX:101453396-101544843	
HC	chr2:159279068-159297701	uc002uab.1,	chr2:159297701-159314628	
HC	chr3:196815575-196964433	uc010hzq.1*,uc003fuz.1*,uc003fva.1*, uc003fvb.1*,uc003fvd.1*,uc003fve.1*, uc003fvc.1*,uc010hzr.1*,uc003fvm.1*, uc003fvg.1*,uc003fvf.1*,uc003fvh.1*, uc010hzs.1*,uc003fvi.1*,uc003fvj.1*, uc003fvk.1*,uc003fvl.1*,uc010hzt.1*, uc003fvo.1*,uc003fvp.1*,	chr3:198829062-198875855	
HC	chr7:2527412-2547651	uc003smh.2*,uc003smi.1*,	chr7:2668424-2688424	
HC	chr11:82722470-82742470	uc001pag.2,	chr11:82795847-82820058	
HC	chr13:41513616-44871138	uc001vaf.2*,uc001vag.1*,uc001vah.1*,	chr13:44895778-45937095	
HC	chr16:2711236-2731236	uc002crh.1*,uc002cri.1*,	chr16:2842485-2872103	
HC	chr16:20410027-20544498	uc002dhm.1*,uc002dhn.1*,uc010bwg.1*,	chr16:20626906-20644419	
HC	chr22:17309884-17383632	uc002zon.1*,	chr22:18663258-19074044	
HCR	chr2:85383400-85403400	uc002soz.1,uc002spa.1,uc002spb.1, uc010fgd.1,	chr2:85407357-85428346	
HCR	chr3:69614222-69632978	uc003dnw.2,	chr3:69632978-69657841	
HCR	chr4:75009154-75111904		chr4:75215771-75235771	uc003hhn.1, uc003hho.1, uc003hhp.1,
HCR	chr10:29746730-29772951	uc001iuo.1*,uc001iup.2*,uc001iuq.1*,	chr10:30901853-31037464	
HCR	chr19:61520299-61593702	uc002qmy.1*,uc002qmz.1*,	chr19:61689010-61703865	
HCR	chr5:99827040-99849238	uc003kni.1,	chr5:99877242-99897242	
HCR	chr18:1130677-1150677		chr18:1340832-1364961	uc002kld.1,
HCR	chr3:116193598-116214464	uc003ebm.1,uc003ebn.1,uc003ebp.2,	chr3:116214463-116235103	

HCR	chr22:19686704-19703212	uc002ztz.2*,uc010gsu.1*,uc002zua.2*, uc002zuc.1*,	chr22:19714784-19744782	
HCR	chr12:108961992-108981992	uc001tqd.1*,uc001tqe.1*,uc001tqf.1*,	chr12:119986740-120006740	
HCR	chr19:62502809-62790802	uc002qpg.1*,uc002qph.1*,uc002qpi.1*, uc002qpj.1*,	chr19:62804675-63134919	
Mouse	chr5:10164215-10191058	uc008wlm.1*	chr5:10231411-10270272	

Supplementary Table 1. A number of mammalian reversals overlap genes that could correspond to lineage-specific innovations. For each such reversal, the coordinates of the two breakpoint regions are shown along with the full list of UCSC genes overlapping each breakpoint. Genes labeled with * overlap both the breakpoint regions.

chromosome	Bp1	Bp2	Published bp1	Published bp2	Lineage
Chr4	44506398-44506432	86178659-86180659	44563003-44603003	86255684-86296684	Chimp
Chr5	18586013-18589073	95946183-95948183	18582661-18622611	95031126-96011126	Chimp
Chr12	20854517-21291427	66657770-66677770	20845308-20885308	66631594-66671594	Chimp
Chr17	7865108-7876865	44965502-44985502	8123673-8163673	48068346–48108346	Chimp

Supplementary Table 2. Coordinates of 4 chimp-specific centromeric inversions mapped on the human genome (hg18). Published breakpoint regions are mapped from hg17 to hg18 using the liftOver program from the UCSC genome browser.

	Human	HC	HCR	Chimp	Rhesus	Mouse	Rat	Dog
Revs < 100Kb	5	15	63	13	38	81	170	165
Revs > 100Kb with SD	6	11	10	NA	NA	NA	NA	NA
Revs > 100Kb without SD	1	2	5	4	11	8	50	15
Total	12	28	78	17	49	89	220	180

Supplementary Table 3. Reversals for which both breakpoints are defined within 100Kb. "Revs < 100Kb" represents the reversals with well-defined breakpoints that will be analyzed for repeat content. "Revs >100 Kb" will be excluded from the repeat analysis because 1) their breakpoints are not precisely determined but also 2) because of the computational limitation in performing simulations. We note that in the primate lineage, SDs supports most of the excluded reversals.

	Human	HC	HCR	Chimp	Rhesus	Rat	Mouse	dog
Predicted reversals	12	29	83	17	49	227	90	184
Actual reversals	12	28	78	17	49	220	89	180
Predicted transpositions	4	15	8	7	40	127	10	17
Actual transpositions	4	14	8	7	40	125	10	17

Supplementary Table 4. List of events that can be localized (i.e. actual). It is easy to see that almost all of the predictions on the leaf edges and in the primate lineages can be localized.

	Number of events				Normalized number of events			
	Rev	Transp	Tloc	Fus/Fis	Rev	Transp	Tloc	Fus/Fis
human-chimp	66	11	3	1	10	1.67	0.45	0.15
human-rhesus	197	61	19	2	10	3.10	0.96	0.10
human-mouse	541	92	138	6	10	1.70	2.55	0.11
human-rat	895	169	155	9	10	1.89	1.73	0.10
human-dog	490	53	76	17	10	1.08	1.55	0.35
mouse-rat	629	134	49	9	10	2.13	0.78	0.14

Supplementary Table 5. The number of rearrangement events and normalized ratios for six pair of genomes at a 10Kb resolution.

Supplementary References

Newman, T.L., E. Tuzun, V.A. Morrison, K.E. Hayden, M. Ventura, S.D. McGrath, M. Rocchi, and E.E. Eichler. 2005. A genome-wide survey of structural variation between human and chimpanzee. *Genome Res* **15:** 1344-1356.

Zhao, H. and G. Bourque. 2007. Recovering True Rearrangement Events on Phylogenetic Trees. In *Comparative Genomics, RECOMB 2007 International Workshop, RECOMB-CG 2007* eds G. Tesler and D. Durand), pp. 149-161. Springer, San Diego, CA, USA.