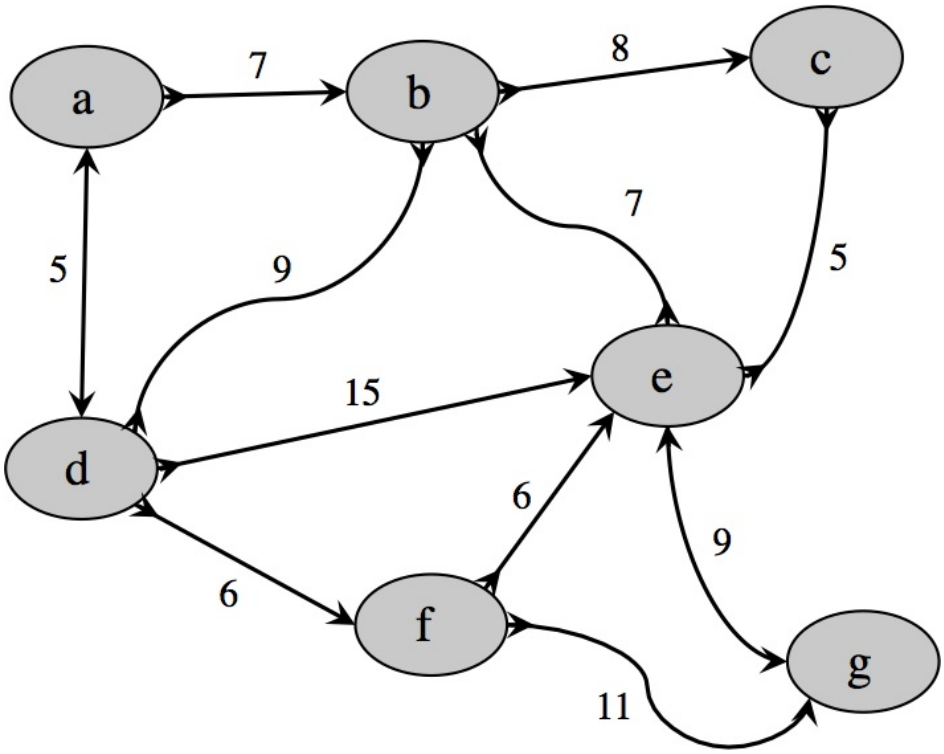
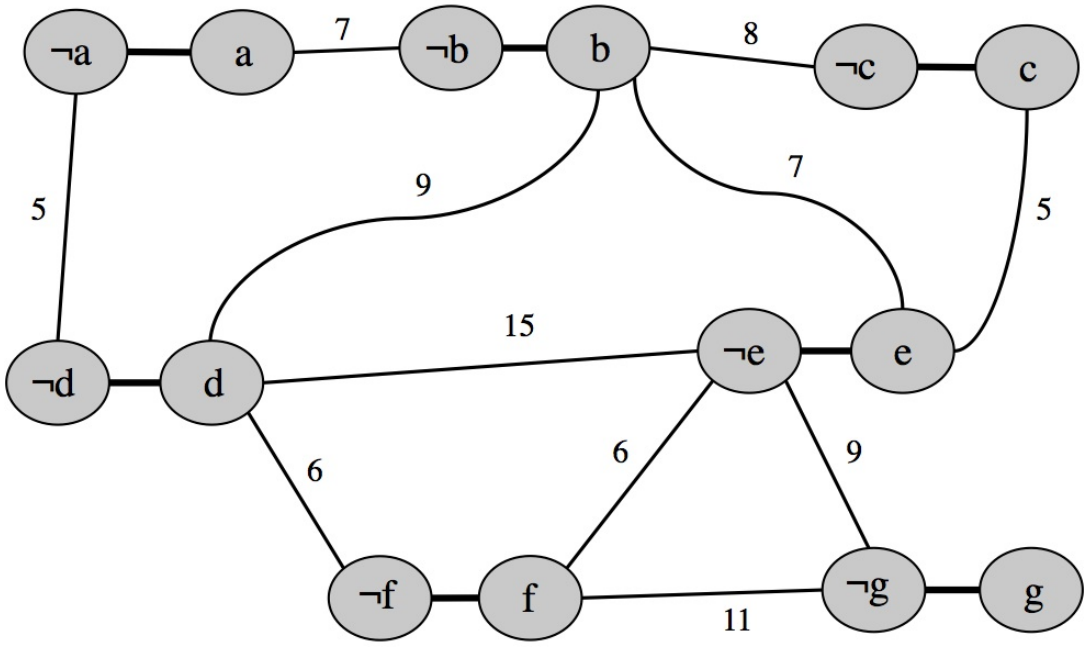
1 MAX-DIR to MAX-CUT Reduction

In the reduction from MAX-DIR to MAX-CUT problems, essentially two nodes $\neg a$ and a are created for each bidirected node a with an "infinitely" weighted edge between them (to ensure that the edge is cut in the MAX-CUT solution). For each bidirected edge e spanning nodes a and b in the MAX-DIR instance, an undirected edge is created in the MAX-CUT instance between either $\neg a$ or a and $\neg b$ and b, depending on the edge-orientations of e with respect to a and b in the MAX-DIR instance (see Fig. 1).



(a) MAX-DIR Instance



(b) MAX-CUT Reduction

Figure 1: (a) An instance of a MAX-DIR problem (note that this example is slightly different than that presented in the text) (b) A reduction to an instance of the MAX-CUT problem.

2 Scaffold Graph Generation Details

2.0.1 Synthetic Genome (w/o Errors)

A 1.25Mb genome with haplotype variation (heterozygosity rate ≈4.0%) was synthesized from chromosome 25 of the zebra finch (ACCN NC_011489.1) using an in-house software package called HapMaker [7]. The following were generated for contig assembly: a set of 250bp single reads at 80x coverage; a 4kb paired-end library at 8x coverage (250bp reads); and a 20kb paired-end library at 8x coverage (250bp reads). Newbler 2.6 was used to assemble contigs with the following parameters: -mi 98 -minlen 30 -infoall -ml 50 -a 1 -nohet -large. Reads were mapped and a scaffold graph formed using ScaffoldScaffolder [1] with Bowtie v0.12.9 [5]. Only the 4kb libraries were used for the scaffold graph construction. ScaffoldScaffolder was run with the parameters -minalign 30 -algorithm greedy -minsupport 1 –showExcludedEdges -ploidy 2. Parameters for Bowtie were -v 2 -a -m 1 -f.

2.0.2 Synthetic Genome (w/ Errors)

Using the 1.25Mb genome reference, a 200bp paired-read library was generated from ART v1.3.1 [2] using a quality profile with an estimated error rate of 3.35%. ART parameters were -l 75 -f 80 -qs 0 -s 10 -m 200. Contig assembly was performed using Newbler, and a scaffold graph was created using ScaffoldScaffolder and Bowtie2 v2.0.6 [4]. Bowtie2 is better designed to handle errors and indels than Bowtie. Newbler was run with parameters -mi 90 -minlen 30 -infoall -ml 50 -a 1 -nohet -large. ScaffoldScaffolder was run with parameters -minalign 30 -algorithm greedy -minsupport 1 –showExcludedEdges -ploidy 2. Bowtie2 parameters were –end-to-end -f -k 1 –score-min L,-0.6,-1.2.

2.0.3 Raspberry Genome

We assessed performance on a scaffold graph constructed from contigs assembled using Newbler on reads from a *Rubus idaeus cultivar heritage* raspberry genome (350 Mb). Contigs for the raspberry genome were assembled using reads from a combination of Illumina HiSeq and 454 sequencing technologies. Single reads and short insert Illumina paired reads were sampled at high coverage and paired 454 reads (with insert sizes ranging from 3kb to 20kb) were sampled at low coverage. Newbler was run with parameters -mi 90 -minlen 30 -infoall -ml 50 -a 1 -nohet -large. Due to the size of the graph, only contigs greater than 500 bp in length were considered in the graph. A scaffold graph was generated using ScaffoldScaffolder and Bowtie2. Only Illumina reads were used for scaffolding inference. ScaffoldScaffolder was run with parameters -minalign 30 -algorithm greedy -minsupport 1 –showExcludedEdges -ploidy 2. Bowtie2 parameters were the same as with the synthetic genome with errors.

2.0.4 Strawberry Genome

Contigs were assembled for *Fragaria vesca* [8] using Newbler on 7 LS454 (454 GS FLX Titanium) runs and 4 LS454 (454 GS FLX) runs (see NCBI Study SRP004241). Newbler was run with parameters -minlen 20 -ml 50 -mi 86 -a 1 -nohet -large. Contigs of length 500bp or greater were used. A scaffold graph was generated from ScaffoldScaffolder and Bowtie2 using paired reads from two 3kb paired-end libraries. Parameters were the same as with the raspberry genome scaffolding and read mapping.

2.0.5 Oyster Genome

Contigs for the Pacific oyster *Crassostrea gigas* were assembled using fosmid pooling in a manner similar to that presented in [9]. A 400bp insert library was generated at 60x for each of 1613 fosmids. These were then sequenced using the Illumina HiSeq 2000 protocol. Each fosmid read pool was assembled using the SOAPdenovo 2.04 [6] *sparse_pregraph* module and parameters -R -K 63 -z 6000000. Resulting contigs were then jointly assembled using the same module and parameters -K 63 -z 900000000. Only contigs of length 500bp or greater were used in the pursuant analysis. A scaffold graph was generated from ScaffoldScaffolder and Bowtie2 using paired reads sequenced from 170bp inserts (Illumina HiSeq 2000 protocol). Reads used for scaffolding were first error corrected using Quake [3] with parameters –no_count -k 19. Parameters were the same as with the raspberry genome scaffolding and read mapping.

2.0.6 Human Genome

Contigs for the HapMap individual 19240 were assembled using SOAPdenovo2 on 2x126bp paired Illumina reads and 2x250 paired Illumina reads (estimated insert size for all reads was 550). SOAPdenovo was run with several values of k to find an optimal assembly (k=87 was selected). These contigs were used with ScaffoldScaffolder and Bowtie to produce a scaffold graph. Only 2x126bp paired Illumina reads were used in scaffolding. Accession numbers for NA19240 reads are ERR899710, ERR899709, ERR894724, ERR894723, ERR899711, and ERR309934. ScaffoldScaffolder was run with parameters -minalign 60 -algorithm greedy -minsupport 2 –showExcludedEdges -ploidy 2. Bowtie was run with -v 3 -a -m 1 –best.

2.0.7 HsInv0393 Region

Reads from ERR899710, ERR899709, ERR894724, ERR894723, ERR899711 were aligned to the hg19 chromosome X reference. All reads aligning to chrX:100829790-100894015 were then assembled using Newbler and SOAPdenovo2. Newbler parameters included -nohet. SOAPdenovo2 was run with k=87 and the resulting scaffolds (not contigs) were used. Newbler contigs and SOAP scaffolds were aligned to hg19 chromosome X reference. All Newbler contigs were used and a single SOAPdenovo contig spanning a region not covered by any Newbler contigs was also used. This set of 5 sequences was then scaffolded using ScaffoldScaffolder and Bowtie. ScaffoldScaffolder was run with parameters -minalign 60 -algorithm greedy -minsupport 2 –showExcludedEdges -ploidy 2. Bowtie was run with parameters -v 3 –best -q.

3 Full Results Table

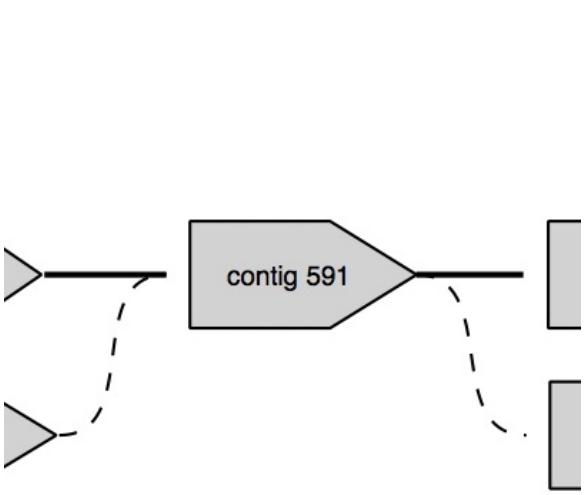
Table 1: Comparative Results

	Greedy	RandEdge	BiqMac	LPSolve	SCIP	GLPK	SDP	Sahni	Random
Synthetic w/o Errors									
Retained Edges									
Count	791	790	748	619	542	541	637	624	407
Weight	6588	6553	6454	5802	5215	5215	5827	5190	3363
Excluded Edges									
Count	2	3	45	174	251	252	156	169	386
Weight	9	44	143	795	1382	1382	770	1407	3234
Synthetic w/ Errors									
Retained Edges									
Count	1689	1695	1543	1149	1060	1060	1157	1379	857
Weight	39507	39172	38226	34121	31533	31533	31210	31566	19553
Excluded Edges									
Count	25	19	171	565	654	654	557	335	857
Weight	225	560	1506	5611	8199	8199	8522	8166	20179
Raspberry									
Retained Edges									
Count	642834	528616	559199	454275	454435	454346	n/a	496709	429205
Weight	9080859	6604235	7501716	5844177	5844177	5844177	n/a	6822112	4797287
Excluded Edges									
Count	215707	329925	299342	404266	404106	404195	n/a	361832	429336
Weight	525426	3002050	2104569	3762108	3762108	3762108	n/a	2784173	4808998
Time (sec)	63628	614	16503	133224	10708	61115	n/a	4	3
Strawberry									
Retained Edges									
Count	122174	115875	n/a	100439	100506	100467	n/a	115188	94207
Weight	350469	273999	n/a	287866	287866	287866	n/a	318109	215752
Excluded Edges									
Count	66107	72406	n/a	87842	87775	87814	n/a	73093	94074
Weight	80041	156511	n/a	142644	142644	142644	n/a	112401	214758
Time (sec)	478	11	n/a	11441	722	3054	n/a	1	1
Oyster									
Retained Edges									
Count	15162	13810	11856	9928	9934	9937	n/a	7358	8367
Weight	315383	273863	295663	242891	242891	242891	n/a	175681	163683
Excluded Edges									
Count	1580	2932	4886	6814	6808	6805	n/a	9384	8375
Weight	2692	44212	22412	75184	75184	75184	n/a	142394	154392
Time (sec)	2	0	780	35	5	17	n/a	2	2

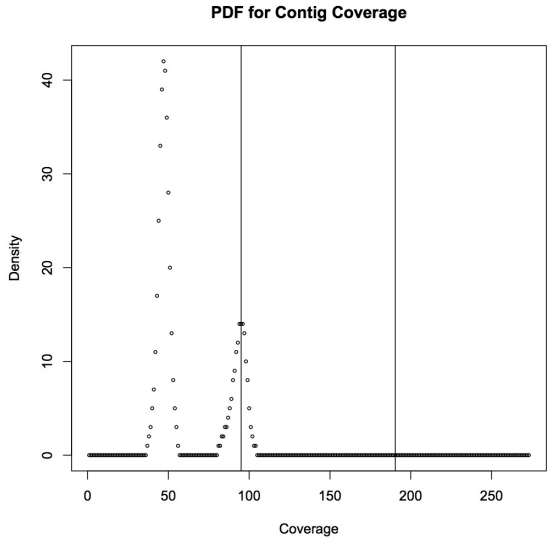
Table 2: Comparative Results (cont.)

	Greedy	RandEdge	BiqMac	LPSolve	SCIP	GLPK	SDP	Sahni	Random
Human									
Retained Edges									
Count	3897	3824	n/a	3144	3144	3145	n/a	2856	2009
Weight	162722	153096	n/a	145553	145553	145553	n/a	125208	83703
Excluded Edges									
Count	108	181	n/a	861	861	860	n/a	1149	1996
Weight	605	10231	n/a	17774	17774	17774	n/a	38119	79624
Time (sec)	1	1	n/a	2	1	1	n/a	1	1

4 Inverted Repeat Detection



(a)



(b)

qseqid	sseq	pident	length	qstart	qend	sstart	send	evalue
contig591	Ref	100.00	532	1	532	737796	737265	0.0
contig591	Ref	100.00	532	1	532	741567	742098	0.0

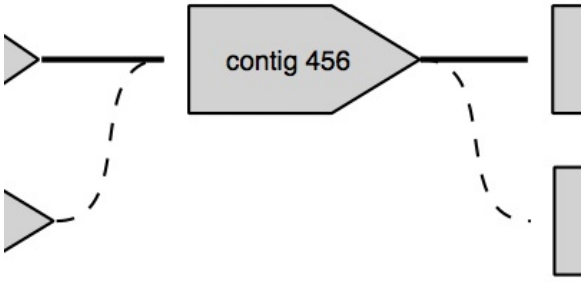
(c)

Figure 2: (a) A closeup of the subgraph produced using the greedy heuristic algorithm on a synthetic genome w/o errors. The only two excluded edges (dotted lines) were both adjacent to contig 591. (b) A probability density function of the sequence depths of all contigs. The 190.4 depth of contig 591 (right vertical line) is almost exactly twice the diploid coverage (left vertical line), suggestive of a repeat contig. (c) The BLAST result of contig 591 to the reference verified that the contig was not only repetitive, but also an inversion.

5 Inverted Haplotype Detection

```
haplotype - Contig 456:
        3' Edges:
                Edge 101-456-: 6 support, 2465.5 average, 71.8
                Edge 456+458-: 3 support, 2862.0 average, 216.
        5' Edges:
                Edge 101-456+: 4 support, 2546.25 average, 158
                Edge 456-458-: 4 support, 2865.0 average, 175.
repeat - Contig 593:
        3' Edges:
                Edge 158-593-: 3 support, 2495.0 average, 69.8
                Edge 18+593-: 1 support, 2761.0 average, 0.0 s
        5' Edges:
                Edge 101-593+: 6 support, 809.6666666666666 av
                Edge 551-593+: 1 support, 3209.0 average, 0.0
repeat - Contig 694:
        3' Edges:
                Edge 7+694-: 2 support, 64.5 average, 50.20458
                Edge 61+694-: 2 support, 1224.5 average, 58.68
        5' Edges:
                Edge 563+694+: 2 support, 3285.5 average, 37.4
                Edge 89+694+: 1 support, 1498.0 average, 0.0 s
```

(a)



(b)

qseqid	*sseqid*	*pident*	*length*	*qstart*	*qend*	*sstart*	*send*	*eval*
contig456-hap	Ref	100.00	838	1	838	743297	744134	0.0
contig456-hap	Var	99.76	838	1	838	744134	743297	0.0
contig593-rep	Ref	100.00	532	1	532	737796	737265	0.0
contig593-rep	Ref	100.00	532	1	532	741567	742098	0.0
contig593-rep	Var	99.62	532	1	532	741567	742098	0.0
contig593-rep	Var	99.25	532	1	532	737796	737265	0.0
contig694-rep	Ref	100.00	375	1	375	811544	811170	0.0
contig694-rep	Ref	99.47	376	1	375	816180	816555	0.0
contig694-rep	Var	100.00	375	1	375	811544	811170	0.0
contig694-rep	Var	99.20	376	1	375	816180	816555	0.0

(c)

Figure 3: (a) The inversion report shows contig 456 whose edges and depth distinguish it as a probable inverted haplotype. (b) A graphic representation of the subgraph at contig 456 showing two excluded edges. (c) The BLAST result of all three inversions from the inversion report, verifying their correct identification and classification. Note that for contigs 593 and 694 the repeats occur in both haplotypes.

References

[1] Paul M Bodily, Jared C Price, Mark J Clement, and Quinn Snell. ScaffoldScaffolder: an aggressive scaffold finishing algorithm. In *Proceedings of the 2012 International Conference on Bioinformatics & Computational Biology*, pages 385–390. CSREA Press, 2012.

[2] Weichun Huang, Leping Li, Jason R Myers, and Gabor T Marth. ART: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, 2012.

[3] David R Kelley, Michael C Schatz, Steven L Salzberg, et al. Quake: quality-aware detection and correction of sequencing errors. *Genome Biology*, 11(11):R116, 2010.

[4] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

[5] Ben Langmead, Cole Trapnell, Mihai Pop, Steven L Salzberg, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[6] Ruibang Luo, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, Yanxiang Chen, Qi Pan, Yunjie Liu, et al. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience*, 1(1):18, 2012.

[7] Nozomu Okuda, Paul M Bodily, Jared C Price, Mark J Clement, and Quinn Snell. HapMaker: synthetic haplotype generator. In *Proceedings of the 2013 International Conference on Bioinformatics & Computational Biology*, pages 370–374. CSREA Press, 2013.

[8] Vladimir Shulaev, Daniel J Sargent, Ross N Crowhurst, Todd C Mockler, Otto Folkerts, Arthur L Delcher, Pankaj Jaiswal, Keithanne Mockaitis, Aaron Liston, Shrinivasrao P Mane, et al. The genome of woodland strawberry (*Fragaria vesca*). *Nature Genetics*, 43(2):109–116, 2010.

[9] Guofan Zhang, Xiaodong Fang, Ximing Guo, Li Li, Ruibang Luo, Fei Xu, Pengcheng Yang, Linlin Zhang, Xiaotong Wang, Haigang Qi, et al. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*, 490(7418):49–54, 2012.