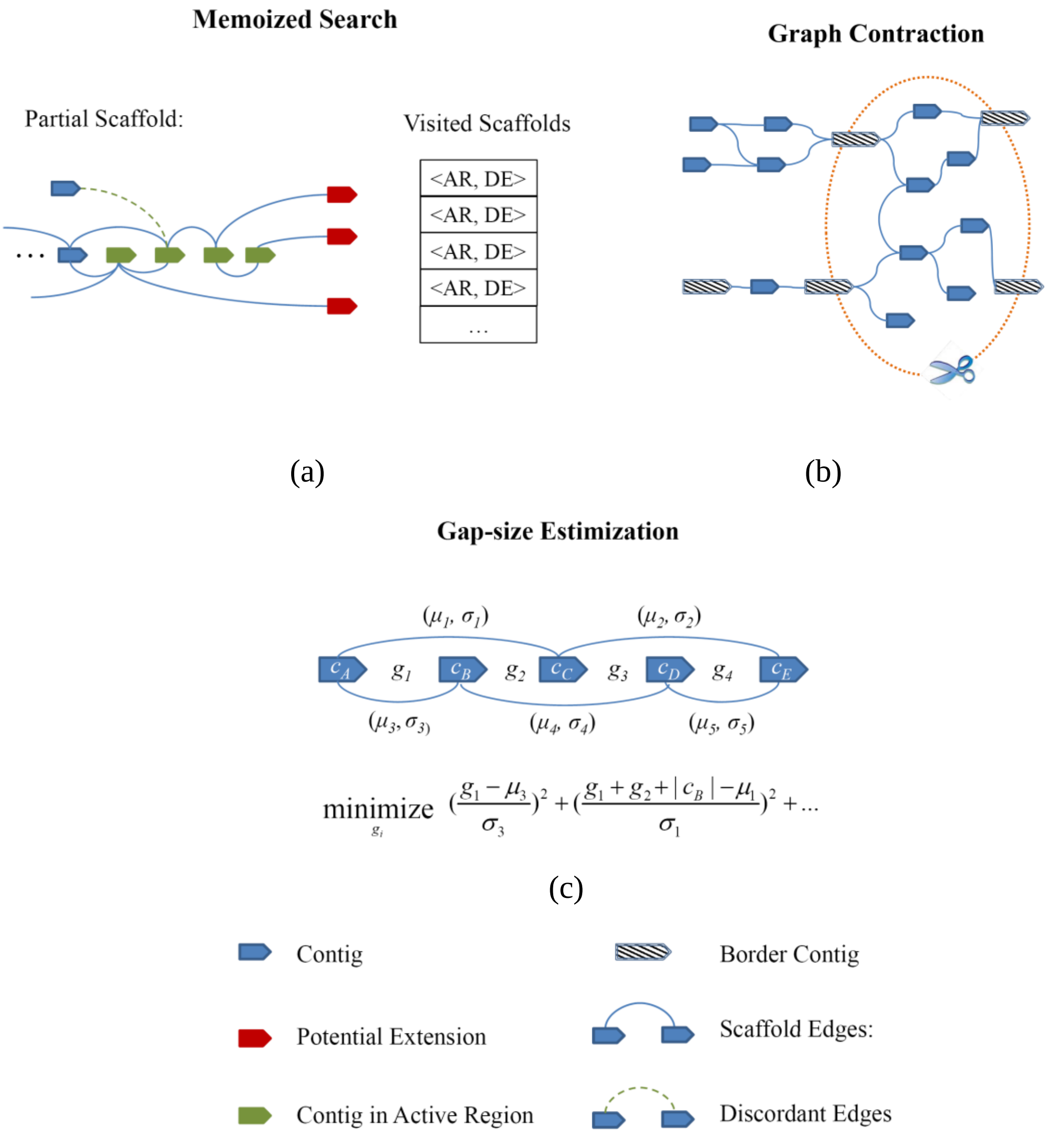
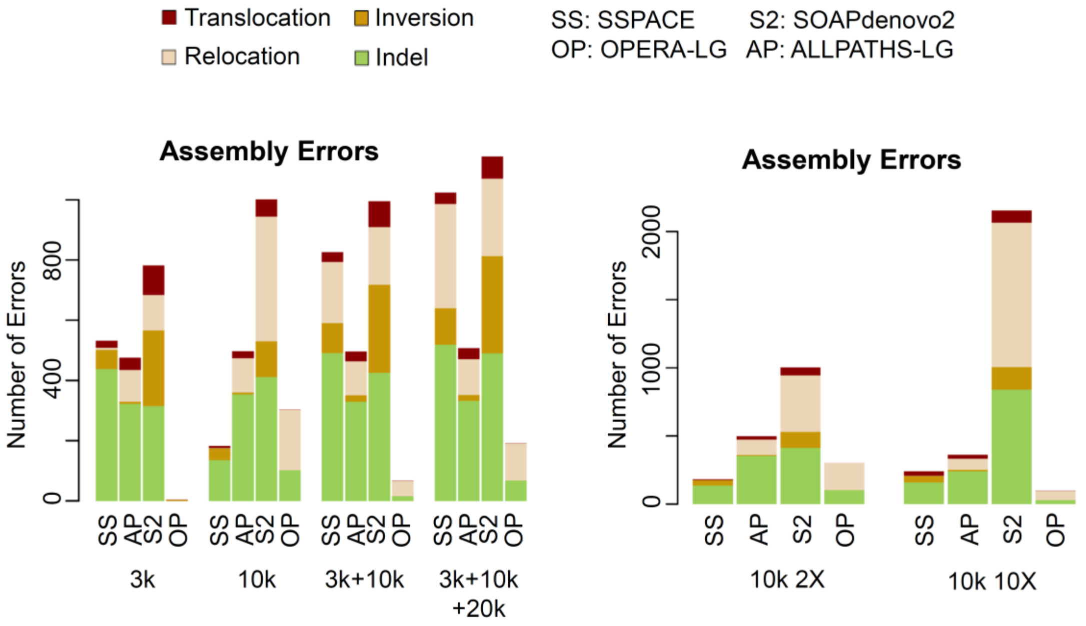
OPERA-LG: Efficient and exact scaffolding of large, repeat-rich eukaryotic genomes with performance guarantees

Song Gao[1*], Denis Bertrand[1*], Niranjan Nagarajan[1]

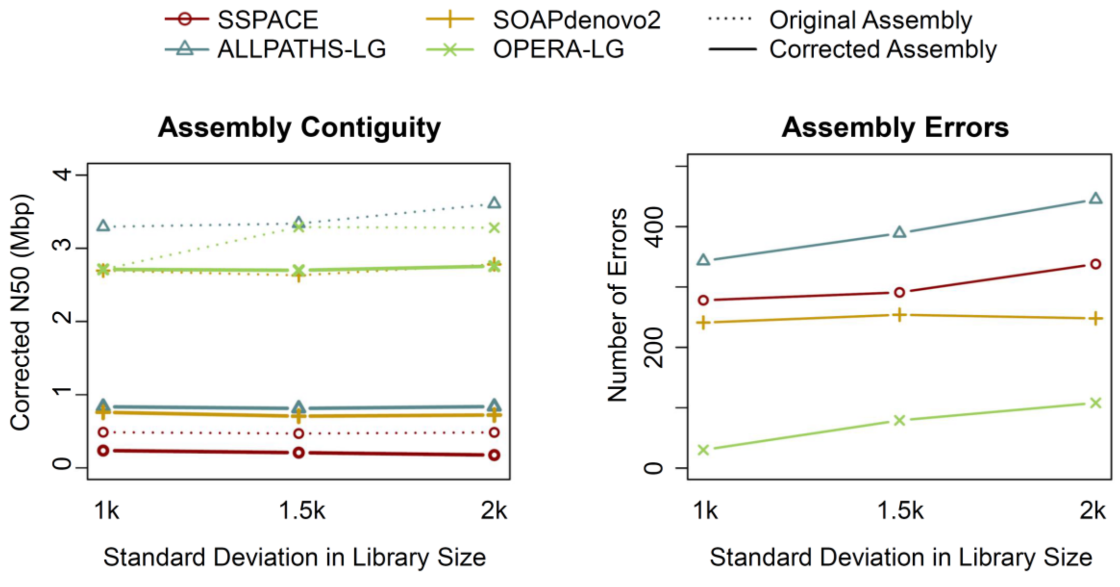


Supplementary Figure 1: **Key algorithmic steps in OPERA-LG.** a) Memoized Search: the search procedure in OPERA-LG is akin to a depth-first search where previously visited partial scaffolds (the tail of which is defined by a list of contigs i.e. “Active Region” or AR and a set of incident edges i.e. “Dangling Edges” or DE) are “memoized” (as <AR, DE> pairs defining an equivalence class of partial scaffolds and not re-searched). b) Graph Contraction: the subgraph demarcated by dotted lines is independently solved in Opera, allowing for significant runtime improvements. Border contigs are large contigs (longer than library size) such that no concordant scaffold edges can span them. c) Gap-size Optimization: gap sizes are jointly optimized in Opera by minimizing the quadratic function depicted in the figure.



(a) (b)

Supplementary Figure 2: **Assembly performance as a function of library information and sequencing depth.** (a) Assembly errors as a function of the mate-pair libraries that were provided as input. (b) Assembly errors as a function of sequencing depth. Results shown here are for the *C. elegans* dataset.



(a) (b)

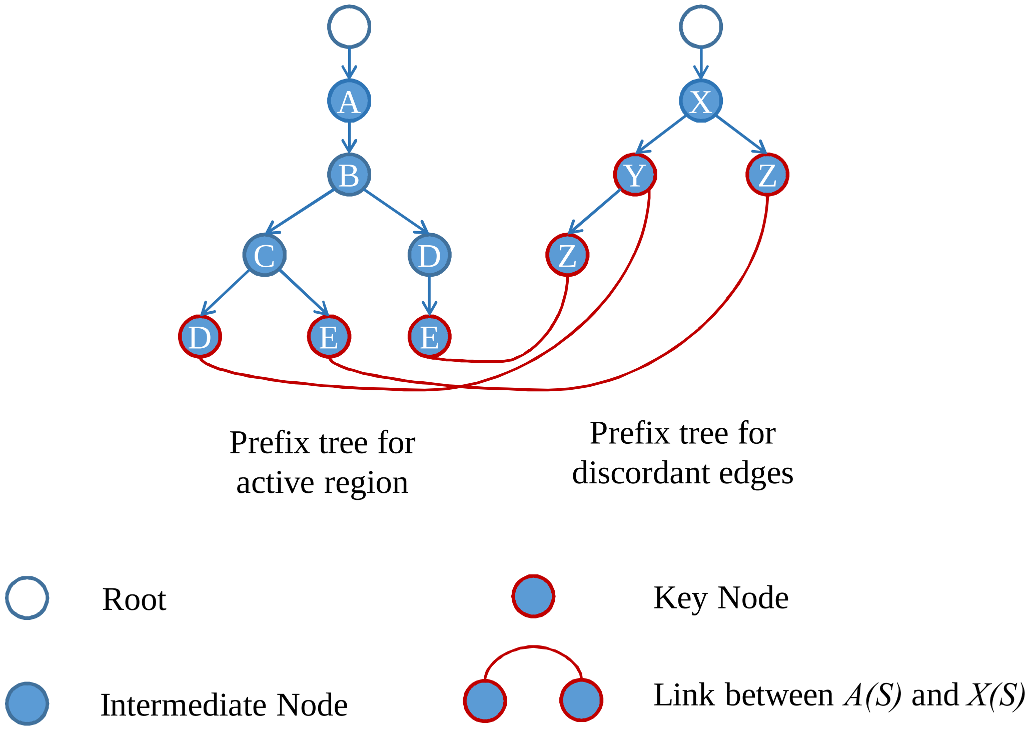
Supplementary Figure 3: **Assembly performance as a function of library quality.** Results shown are for the *D. melanogaster* dataset using 10 kbp libraries.

```
ScaffoldWithRepeat($S', p$)

Require: A scaffold graph $G = (V, E)$ and a partial scaffold $S'$ with at most $p$ discordant edges.

Ensure: Return a scaffold $S$ of $G$ with at most $p$ discordant edges and where $S'$ is a prefix of $S$

if $S'$ is a scaffold of $G$, then
    return $S'$
end if
for every $c \in V - V_{S'}$ in each orientation do
    Let $S''$ be the scaffold formed by concatenating $S'$ and $c$;
    If a confirmed repeat $r$ should be removed then
        trace back to the contig before $r$;
    else
        Let $A$ be the active region of $S''$;
        Let $D$ be the set of dangling edges of $S''$;
        Let $k$ be the number of discordant edges in $S''$;
        if $(A, D, k)$ is unmarked, then
            Mark $(A, D, k)$ as processed;
            if $k \leq p$, then
                $S''' \leftarrow$ ScaffoldWithRepeat($S'', p$);
                if $S''' \neq$ FAILURE, return $S'''$;
            end if
        end if
    end if
end for
Return FAILURE;
```

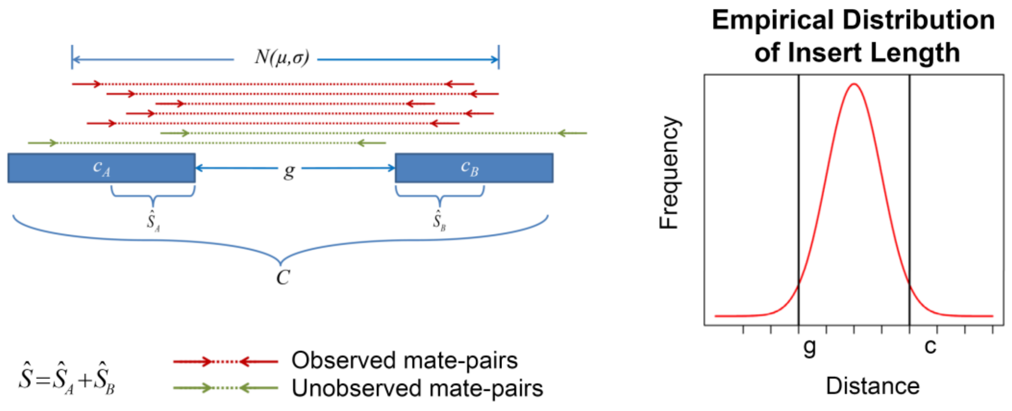
Supplementary Figure 4: An algorithm for generating a minimal-repeat optimal scaffold with at most p discordant edges.

Partial Scaffolds:

S_1: A_1=(A B C D), X_1=(X Y)
S_2: A_2=(A B C E), X_2=(X Z)
S_3: A_3=(A B D E), X_3=(X Y Z)


A
B
C
D
D
E
E
X
Y
Z
Z
Prefix tree for active region
Prefix tree for discordant edges
Root
Key Node
Intermediate Node
Link between *A(S)* and *X(S)*

Supplementary Figure 5: **An example of the prefix tree data structure used to record visited partial scaffolds.** The example here records the partial scaffolds S_1, S_2 and S_3 shown at the top of the figure, where A_i and X_i represent the active region (list of contigs in the tail of the scaffold) and discordant edges, respectively, of the partial scaffolds.



(a) (b)

Supplementary Figure 6: **Observed and un-observed read-pairs.** (a) Graphical depiction of the phenomena of mate-pairs connecting contigs coming from a truncated distribution defined by contig lengths (c_A and c_B) and gap size (g). (b) Empirical distribution of the distance between observed mate-pairs (mean μ and standard deviation σ) and region of truncation (defined by g and C).

		Contigs	Scaffolds
D. melanogaster	SSPACE	87	92
	SOAPdenovo2	86	95
	OPERA-LG	▨	92
	ALLPATHS-LG	89	94
C. elegans	SSPACE	88	107
	SOAPdenovo2	86	102
	OPERA-LG	▨	100
	ALLPATHS-LG	94	100
H. sapiens	SSPACE	71	112
	SOAPdenovo2	66	94
	OPERA-LG	▨	106
	ALLPATHS-LG	79	93

Supplementary Table 1. Assembly size for results reported in Figure 3a-d. The numbers presented here show the total length of contigs and scaffolds (longer than 500 bp) in each assembly, reported as a percentage of genome length. Note that scaffold lengths include gaps and can exceed 100% due to the use of a lower bound for gap sizes in many scaffolders.

		Indel	Inversion	Relocation	Translocation
D. melanogaster	SSPACE	376	33	87	40
	SOAPdenovo2	196	106	104	30
	OPERA-LG	30	0	29	1
C. elegans	SSPACE	854	126	630	39
	SOAPdenovo2	493	323	253	71
	OPERA-LG	67	0	123	0
H. sapiens	SSPACE	9600	1209	8504	676
	SOAPdenovo2	26408	1033	24442	1430
	OPERA-LG	7297	60	4758	94

Supplementary Table 2. Number of scaffold errors as depicted in Figure 3b.

		Indel	**Inversion**	**Relocation**	**Translocation**
D. melanogaster	ALLPATHS-LG	208	38	84	49
	SOAPdenovo2	194	107	101	31
	OPERA-LG	34	0	29	1
C. elegans	ALLPATHS-LG	332	19	119	37
	SOAPdenovo2	490	322	258	73
	OPERA-LG	67	0	123	1
H. sapiens	ALLPATHS-LG	18652	243	5925	3216
	SOAPdenovo2	26833	1187	24485	1754
	OPERA-LG	7310	63	4761	161

Supplementary Table 3. Number of assembly errors (including contig and scaffold errors) as depicted in Figure 3d.

	N50 (Mbp)	**Corrected N50 (Mbp)**	**# of errors**
SSPACE	1.4	0.3	555
SOAPdenovo2	7.0	0.8	574
ALLPATHS-LG	12.0	1.1	432
OPERA-LG	12.0	12.0	14

Supplementary Table 4. Impact of long reads on assembly results. The results reported here are based on redoing the analysis reported in **Figure 3c** for *D. melanogaster*, where 250 bp reads were simulated (instead of 80 bp and genome coverage was kept the same) for the paired-end read library (fragment size 400 bp).