

Enhancing variant detection in complex genomes: leveraging linked reads for robust SNP, Indel, and structural variant analysis

Luo et al.

Contents

1	Supplementary Tables	2
S1	Simulation Configurations	2
2	Supplementary Figures	3
S1	Impact of C_F and C_R combinations on SNP and INDEL calling using single libraries	3
S2	Hybrid PE100_NGS and PE100_stLFR SNP Recall	4
S3	Hybrid PE100_NGS and PE100_stLFR SNP Precision	5
S4	Hybrid SE500_stLFR and PE100_stLFR SNP Recall	6
S5	Hybrid SE500_stLFR and PE100_stLFR SNP Precision	7
S6	Hybrid SE1000_stLFR and PE100_stLFR SNP Recall	8
S7	Hybrid SE1000_stLFR and PE100_stLFR SNP Precision	9
S8	Hybrid PE100_NGS and PE100_stLFR INDEL Recall	10
S9	Hybrid PE100_NGS and PE100_stLFR INDEL Precision	11
S10	Hybrid PE100_NGS and PE100_stLFR INDEL F1	12
S11	Hybrid SE500_stLFR and PE100_stLFR INDEL Recall	13
S12	Hybrid SE500_stLFR and PE100_stLFR INDEL Precision	14
S13	Hybrid SE500_stLFR and PE100_stLFR INDEL F1	15
S14	Hybrid SE1000_stLFR and PE100_stLFR INDEL Recall	16
S15	Hybrid SE1000_stLFR and PE100_stLFR INDEL Precision	17
S16	Hybrid SE1000_stLFR and PE100_stLFR INDEL F1	18
S17	Simulated and real PE100_stLFR SNP INDEL calling perfor- mance comparison	19
S18	Simulated and real PE100_stLFR SV calling performance com- parison	20
S19	Simulated and real PE100_NGS SNP INDEL calling performance comparison	21

1 Supplementary Tables

Shared Parameters: $N_{FL}=1$; Insert Size=600bp; Error Rate=1%						
	PE100/SE500/SE1000 35x			PE100/SE500/SE1000 70x		
	C_F	C_R	$\mu_{FL}(\text{kb})$	C_F	C_R	$\mu_{FL}(\text{kb})$
Exp1	350	0.1	50	700	0.1	50
Exp2	140	0.25	50	280	0.25	50
Exp3	70	0.5	50	140	0.5	50
Exp4	47	0.75	50	94	0.75	50
Exp5	350	0.1	75	700	0.1	75
Exp6	140	0.25	75	280	0.25	75
Exp7	70	0.5	75	140	0.5	75
Exp8	47	0.75	75	94	0.75	75
Exp9	350	0.1	100	700	0.1	100
Exp10	140	0.25	100	280	0.25	100
Exp11	70	0.5	100	140	0.5	100
Exp12	47	0.75	100	94	0.75	100

Table S1: The table summarizes the simulation settings for all experiments (datasets). All datasets share the same N_{FL} , insert size, and error rate. PE100, SE500, and SE1000 indicate the read length and whether sequencing is paired-end or single-end. Experiments marked as 35x or 70x correspond to sequencing depths of 35x or 70x, respectively. Experiments (Exp) 1-12 are divided into three groups: 1-4, 5-8, and 9-12. Across these groups, μ_{FL} varies, while C_F and C_R vary within each group.

2 Supplementary Figures

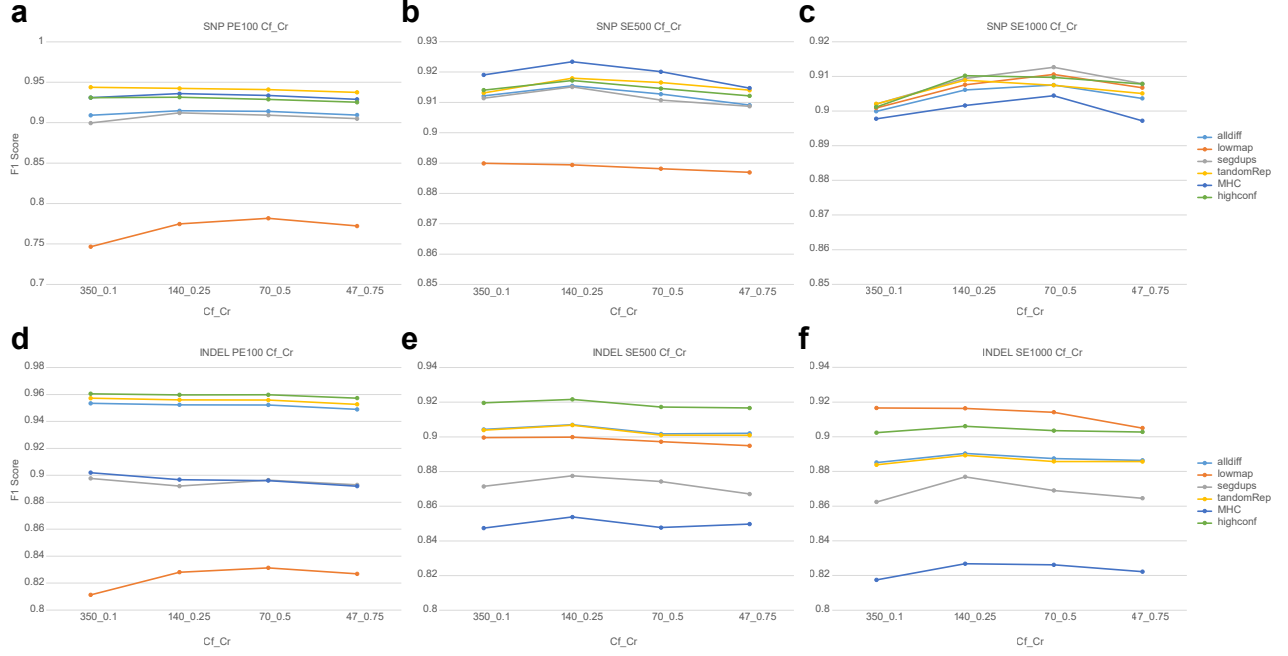


Figure S1: **Impact of C_F and C_R combinations on SNP and INDEL calling using single library.** Impact of C_F and C_R combination was evaluated by varying C_R from 0.1 to 0.75 at a fixed coverage of 35x. **a-c.** F1 scores for SNP calling using PE100, SE500 and SE1000 under various C_F and C_R combinations. **d-f.** F1 scores for INDEL calling using PE100, SE500 and SE1000 under various C_F and C_R combinations.

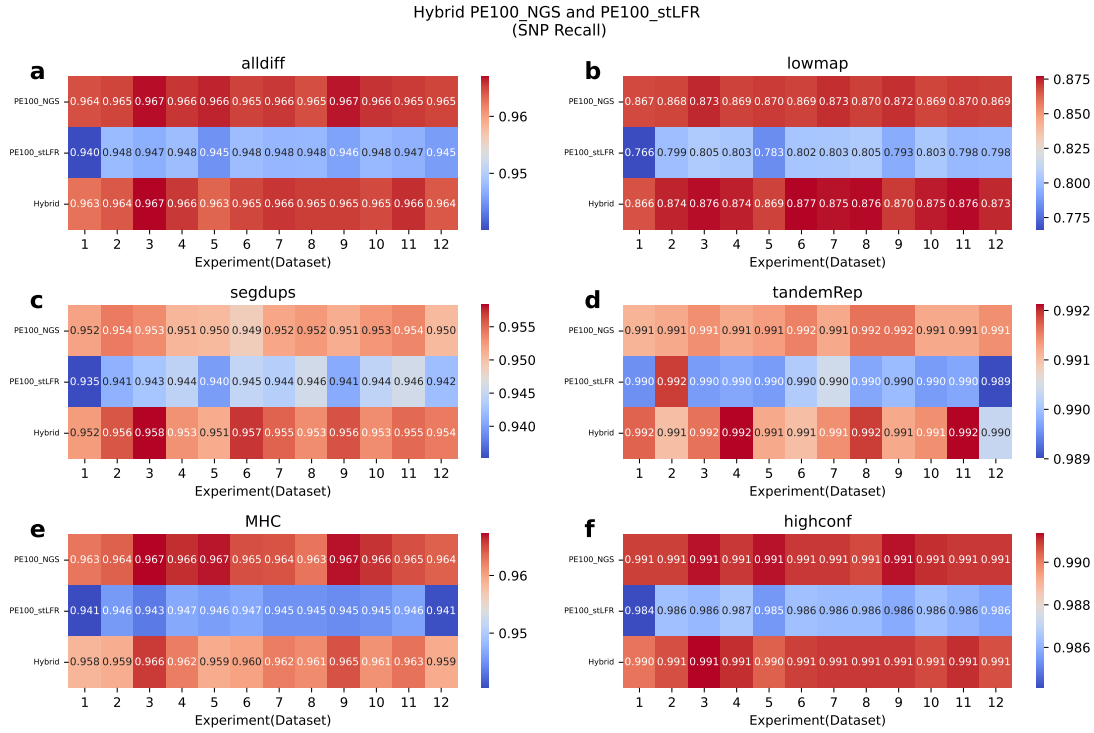


Figure S2: Recall heatmap for SNP calling on hybrid libraries by combining PE100_NGS and PE100_stLFR reads. *hap.py* v0.3.15 was applied to benchmark the SNP callset against the GIAB ground truth callset. **a.** Recall for all difficult (alldiff) regions. **b.** Recall for low-mappability (lowmap) regions. **c.** Recall for segmentation duplication (segdup) regions. **d.** Recall for tandem and homopolymer repeat (tandemRep) regions. **e.** Recall for MHC regions. **f.** Recall for all high-confidence (highconf) regions defined by GIAB.

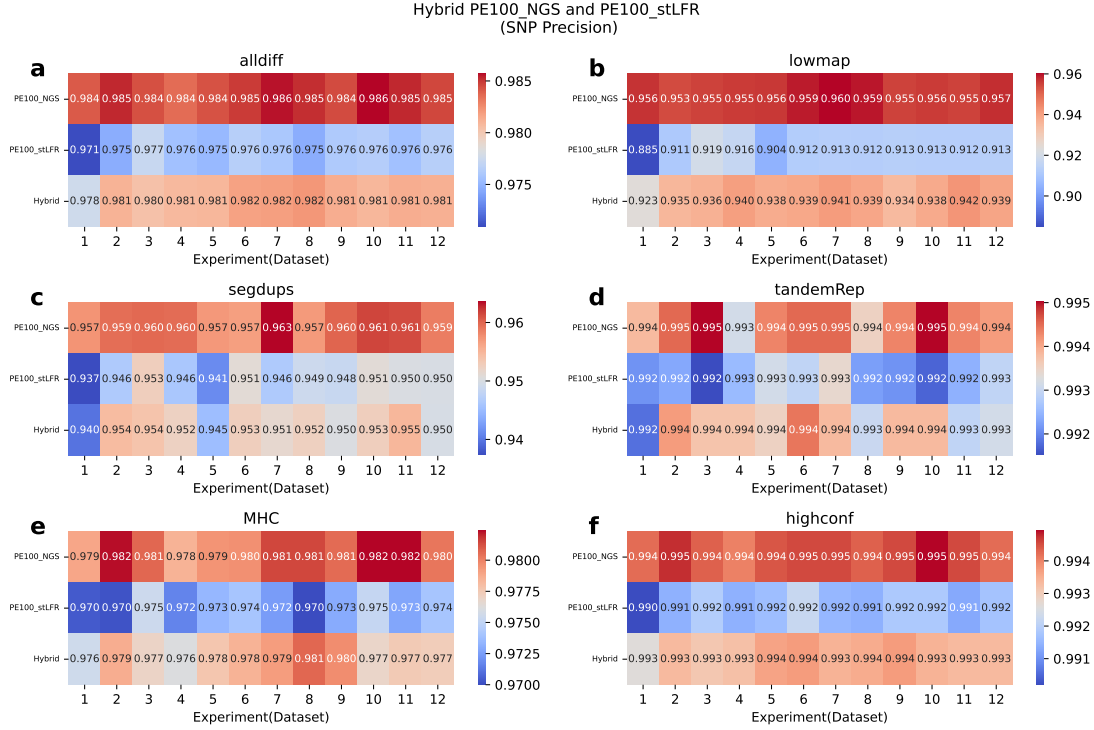


Figure S3: Precision heatmap for SNP calling on hybrid libraries by combining PE100_NGS and PE100_stLFR reads. *hap.py* v0.3.15 was applied to benchmark the SNP callset against the GIAB ground truth callset. **a.** Precision for all difficult (alldiff) regions. **b.** Precision for low-mappability (lowmap) regions. **c.** Precision for segmentation duplication (segdup) regions. **d.** Precision for tandem and homopolymer repeat (tandemRep) regions. **e.** Precision for MHC regions. **f.** Precision for all high-confidence (highconf) regions defined by GIAB.

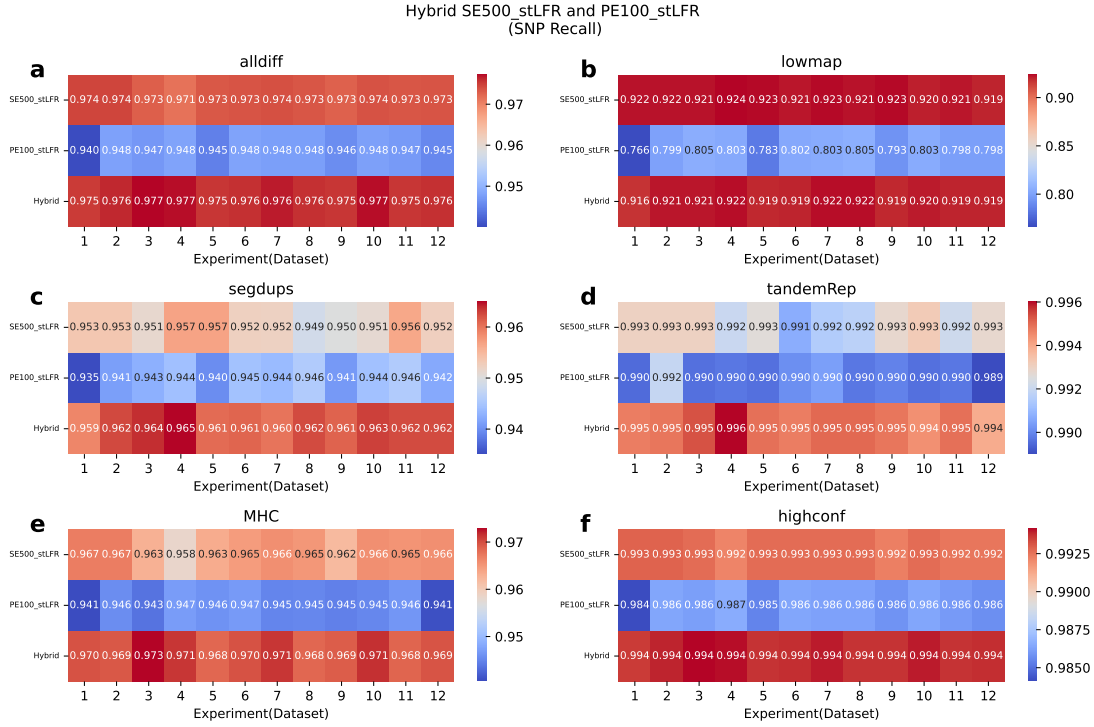


Figure S4: Recall heatmap for SNP calling on hybrid libraries by combining SE500_stLFR and PE100_stLFR reads. *hap.py* v0.3.15 was applied to benchmark the SNP callset against the GIAB ground truth callset. **a.** Recall for all difficult (alldiff) regions. **b.** Recall for low-mappability (lowmap) regions. **c.** Recall for segmentation duplication (segdup) regions. **d.** Recall for tandem and homopolymer repeat (tandemRep) regions. **e.** Recall for MHC regions. **f.** Recall for all high-confidence (highconf) regions defined by GIAB.

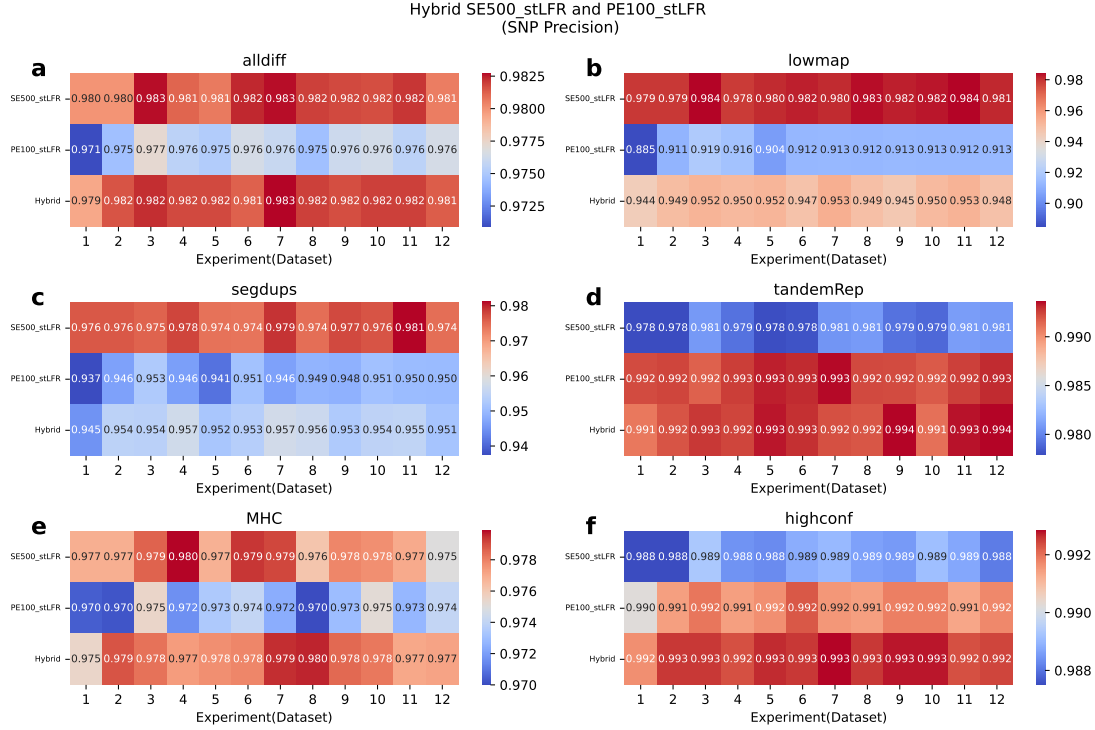


Figure S5: Precision heatmap for SNP calling on hybrid libraries by combing SE500_stLFR and PE100_stLFR reads. *hap.py* v0.3.15 was applied to benchmark the SNP callset against the GIAB ground truth callset. **a.** Precision for all difficult (alldiff) regions. **b.** Precision for low-mappability (lowmap) regions. **c.** Precision for segmentation duplication (segdup) regions. **d.** Precision for tandem and homopolymer repeat (tandemRep) regions. **e.** Precision for MHC regions. **f.** Precision for all high-confidence (highconf) regions defined by GIAB.

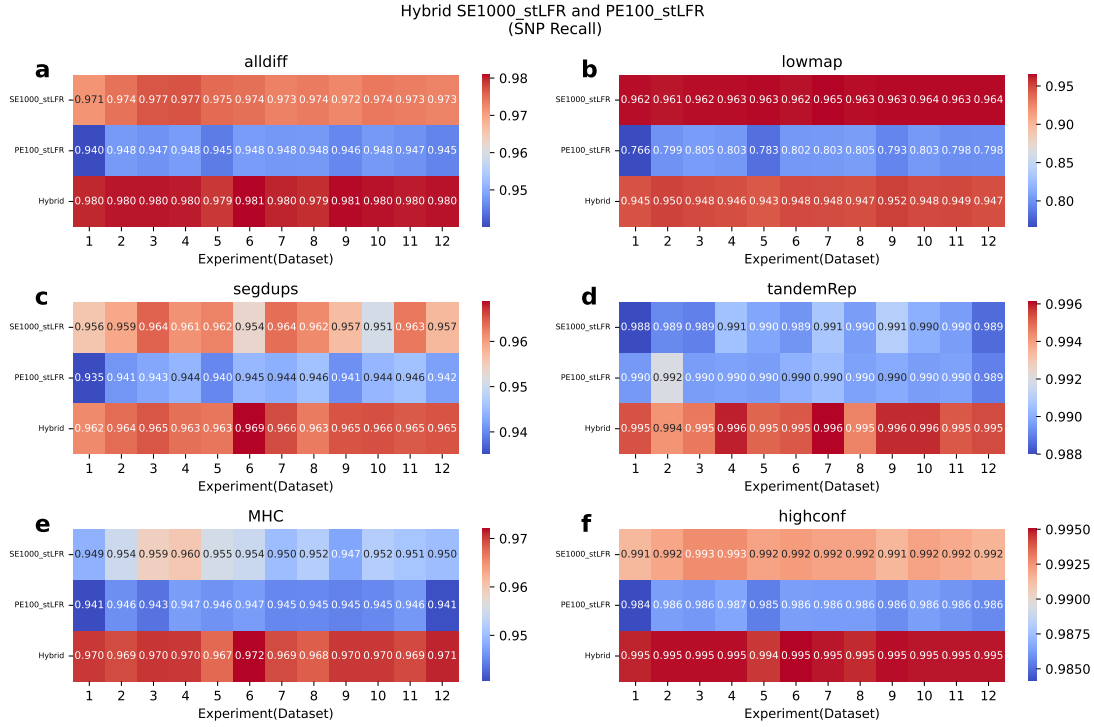


Figure S6: Recall heatmap for SNP calling on hybrid libraries by combining SE1000_stLFR and PE100_stLFR reads. *hap.py* v0.3.15 was applied to benchmark the SNP callset against the GIAB ground truth callset. **a.** Recall for all difficult (alldiff) regions. **b.** Recall for low-mappability (lowmap) regions. **c.** Recall for segmentation duplication (segdup) regions. **d.** Recall for tandem and homopolymer repeat (tandemRep) regions. **e.** Recall for MHC regions. **f.** Recall for all high-confidence (highconf) regions defined by GIAB.

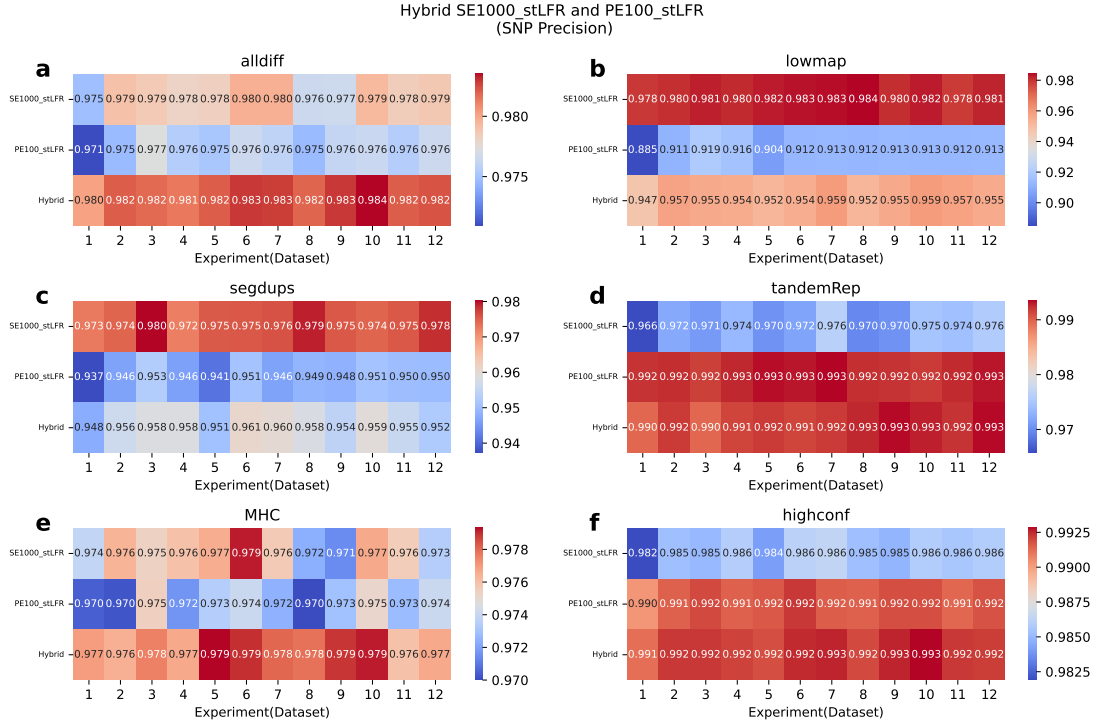


Figure S7: Precision heatmap for SNP calling on hybrid libraries by combining SE1000_stLFR and PE100_stLFR reads. *hap.py* v0.3.15 was applied to benchmark the SNP callset against the GIAB ground truth callset. **a.** Precision for all difficult (alldiff) regions. **b.** Precision for low-mappability (lowmap) regions. **c.** Precision for segmentation duplication (segdup) regions. **d.** Precision for tandem and homopolymer repeat (tandemRep) regions. **e.** Precision for MHC regions. **f.** Precision for all high-confidence (highconf) regions defined by GIAB.

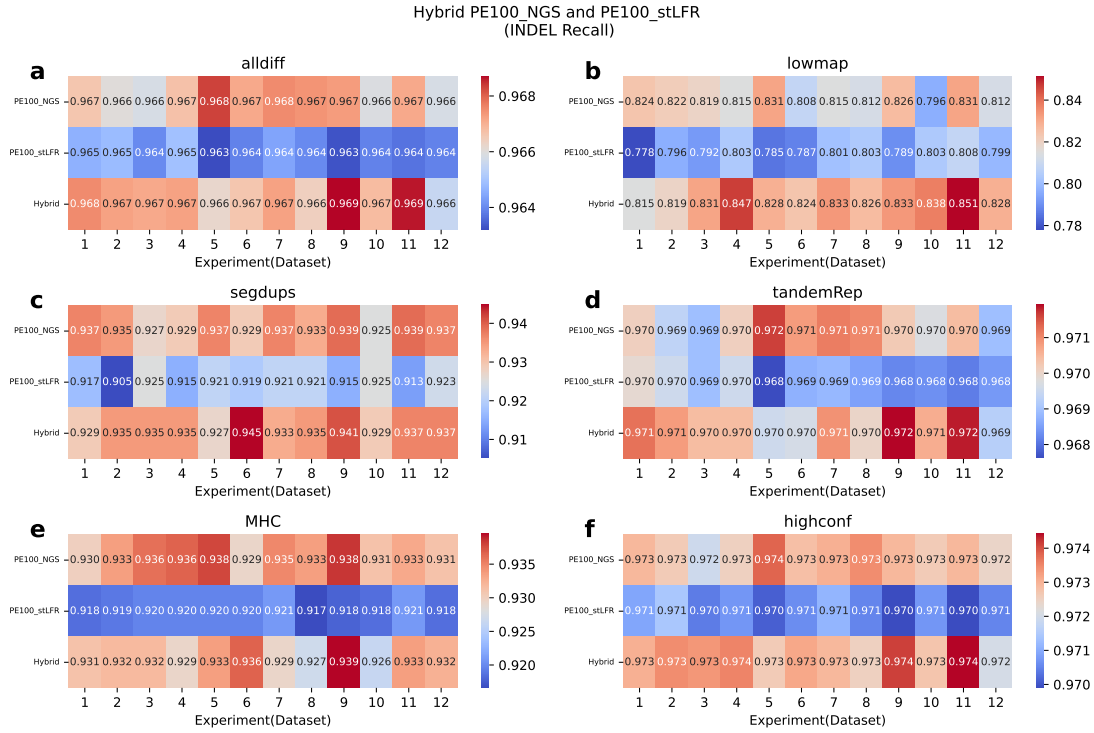


Figure S8: Recall heatmap for INDEL calling on hybrid libraries by combining PE100_NGS and PE100_stLFR reads. *hap.py* v0.3.15 was applied to benchmark the INDEL callset against the GIAB ground truth callset. **a.** Recall for all difficult (alldiff) regions. **b.** Recall for low-mappability (lowmap) regions. **c.** Recall for segmentation duplication (segdup) regions. **d.** Recall for tandem and homopolymer repeat (tandemRep) regions. **e.** Recall for MHC regions. **f.** Recall for all high-confidence (highconf) regions defined by GIAB.

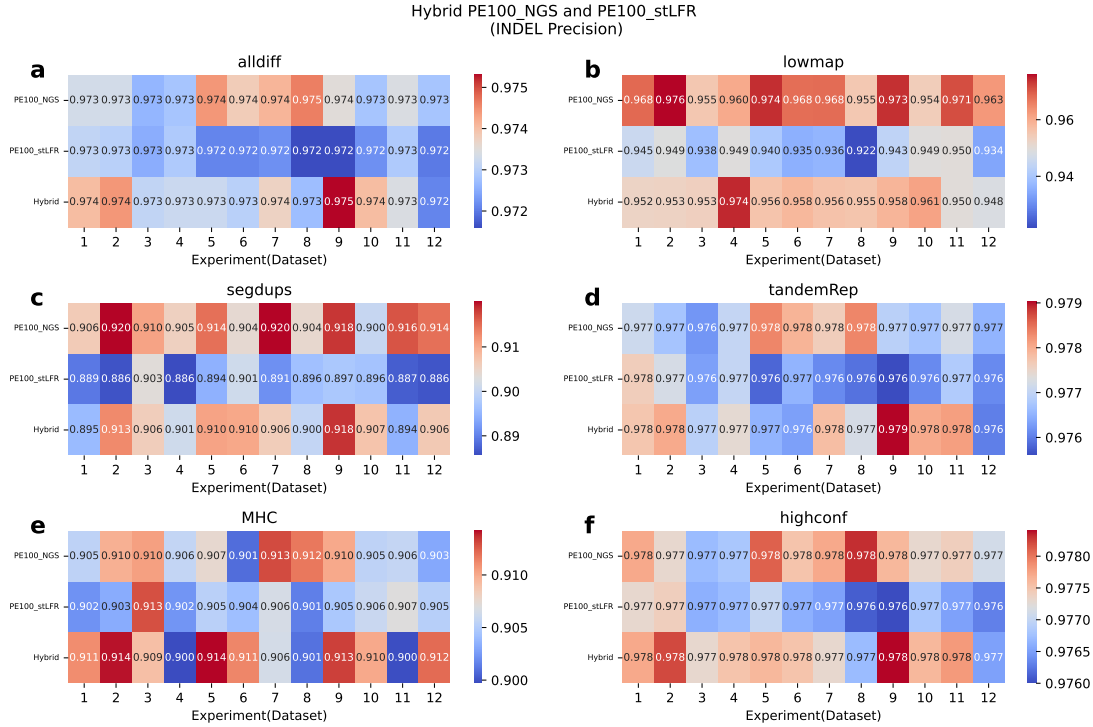


Figure S9: Precision heatmap for INDEL calling on hybrid libraries by combing PE100_NGS and PE100_stLFR reads. *hap.py* v0.3.15 was applied to benchmark the INDEL callset against the GIAB ground truth callset. **a.** Precision for all difficult (alldiff) regions. **b.** Precision for low-mappability (lowmap) regions. **c.** Precision for segmentation duplication (segdup) regions. **d.** Precision for tandem and homopolymer repeat (tandemRep) regions. **e.** Precision for MHC regions. **f.** Precision for all high-confidence (highconf) regions defined by GIAB.

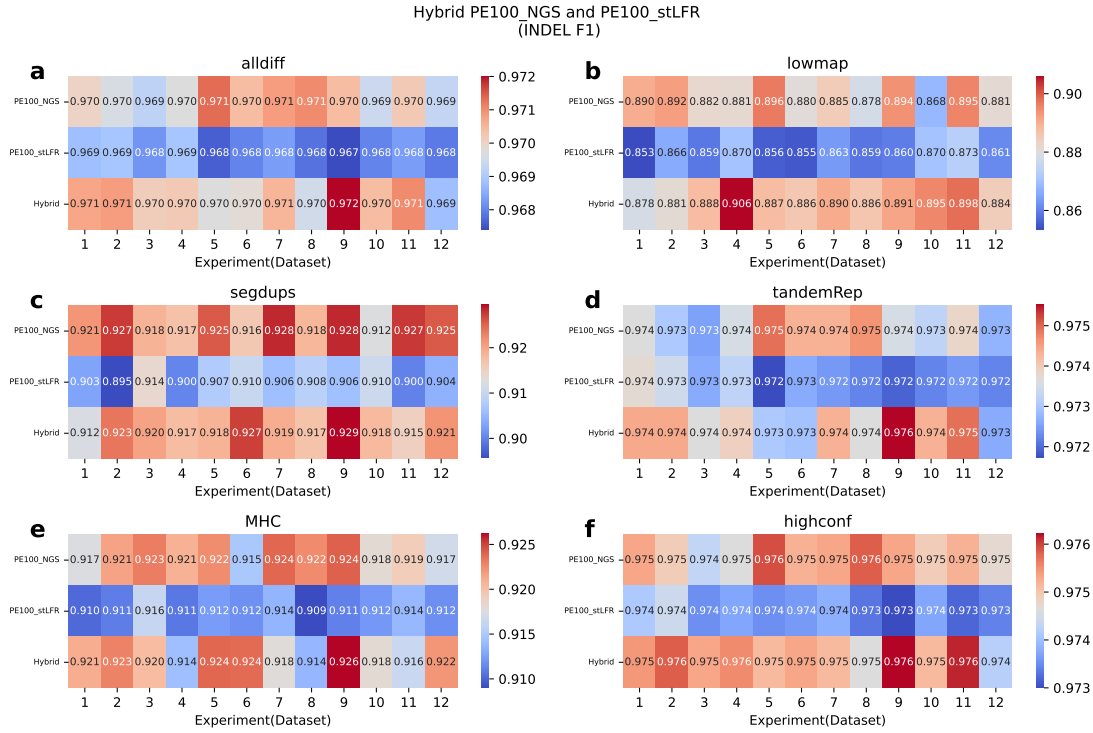


Figure S10: **F1 heatmap for INDEL calling on hybrid libraries by combining PE100_NGS and PE100_stLFR reads.** *hap.py* v0.3.15 was applied to benchmark the INDEL callset against the GIAB ground truth callset. **a.** F1 for all difficult (alldiff) regions. **b.** F1 for low-mappability (lowmap) regions. **c.** F1 for segmentation duplication (segdup) regions. **d.** F1 for tandem and homopolymer repeat (tandemRep) regions. **e.** F1 for MHC regions. **f.** F1 for all high-confidence (highconf) regions defined by GIAB.

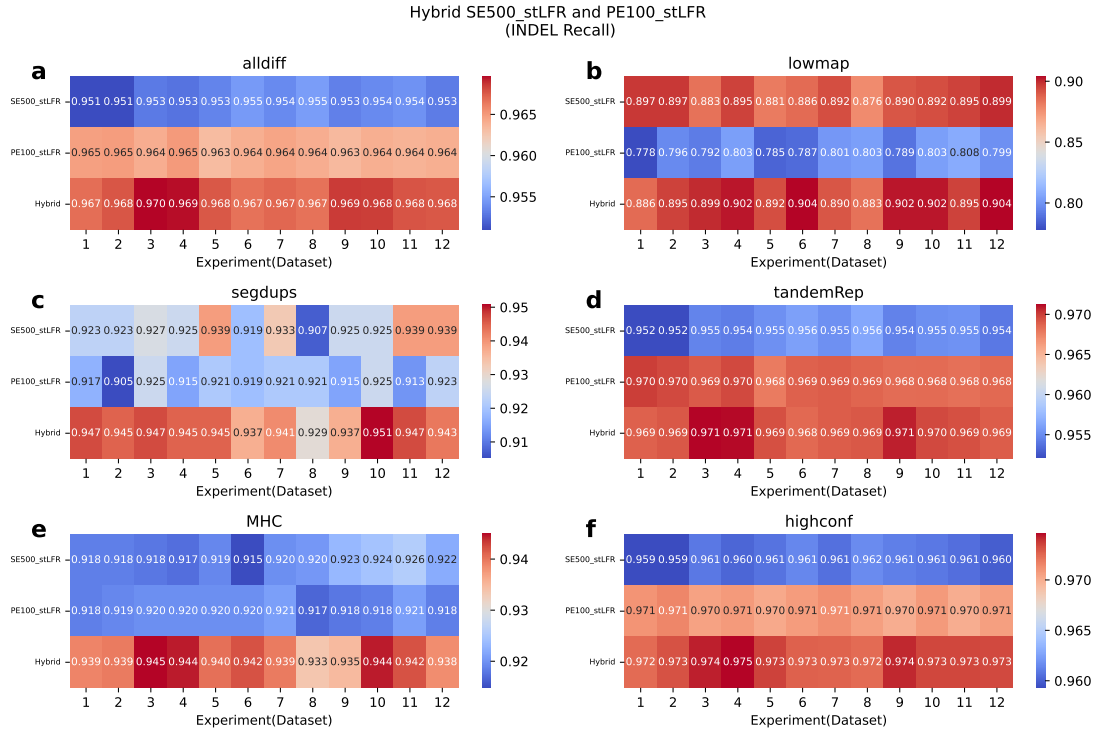


Figure S11: Recall heatmap for INDEL calling on hybrid libraries by combining SE500_stLFR and PE100_stLFR reads. *hap.py* v0.3.15 was applied to benchmark the INDEL callset against the GIAB ground truth callset. **a.** Recall for all difficult (alldiff) regions. **b.** Recall for low-mappability (lowmap) regions. **c.** Recall for segmentation duplication (segdup) regions. **d.** Recall for tandem and homopolymer repeat (tandemRep) regions. **e.** Recall for MHC regions. **f.** Recall for all high-confidence (highconf) regions defined by GIAB.

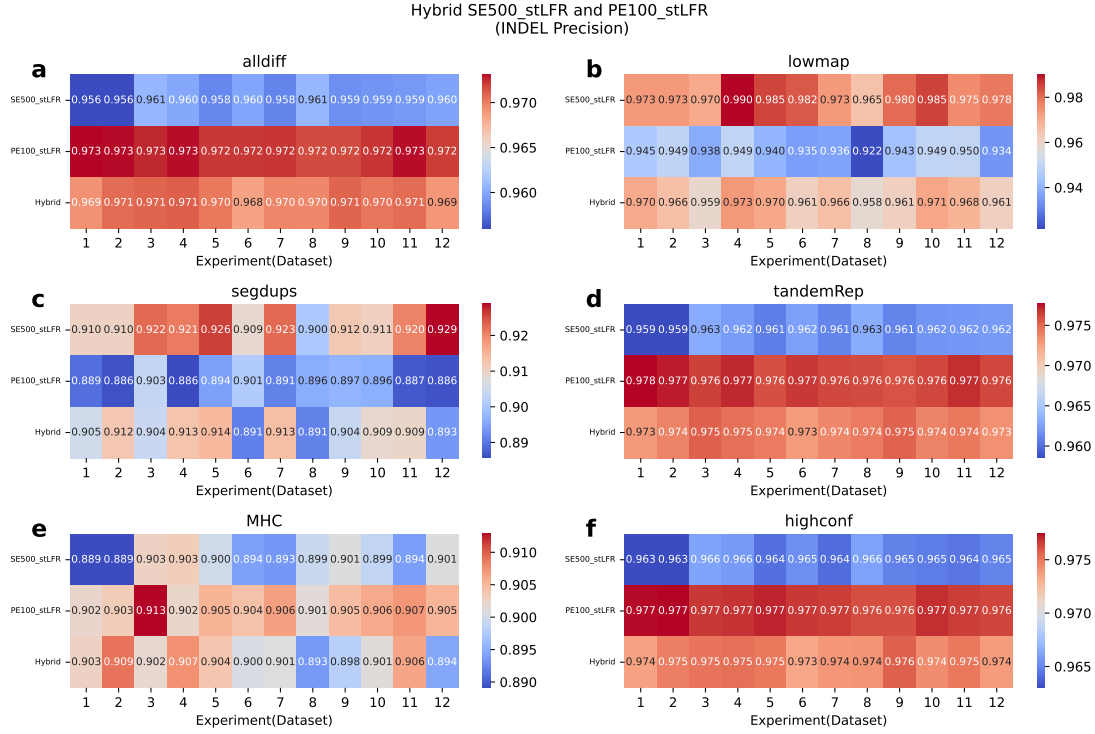


Figure S12: **Precision heatmap for INDEL calling on hybrid libraries by combing SE500_stLFR and PE100_stLFR reads.** *hap.py* v0.3.15 was applied to benchmark the INDEL callset against the GIAB ground truth callset. **a.** Precision for all difficult (alldiff) regions. **b.** Precision for low-mappability (lowmap) regions. **c.** Precision for segmentation duplication (segdup) regions. **d.** Precision for tandem and homopolymer repeat (tandemRep) regions. **e.** Precision for MHC regions. **f.** Precision for all high-confidence (highconf) regions defined by GIAB.

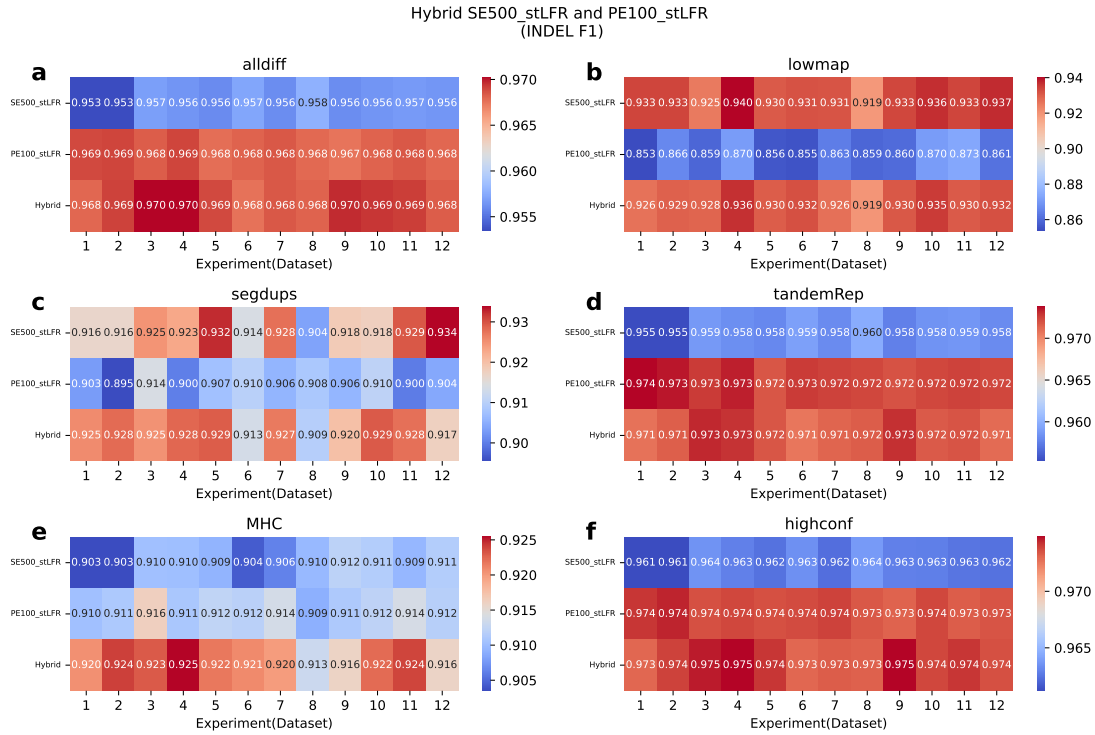


Figure S13: **F1 heatmap for INDEL calling on hybrid libraries by combining SE500_stLFR and PE100_stLFR reads.** *hap.py* v0.3.15 was applied to benchmark the INDEL callset against the GIAB ground truth callset. **a.** F1 for all difficult (alldiff) regions. **b.** F1 for low-mappability (lowmap) regions. **c.** F1 for segmentation duplication (segdup) regions. **d.** F1 for tandem and homopolymer repeat (tandemRep) regions. **e.** F1 for MHC regions. **f.** F1 for all high-confidence (highconf) regions defined by GIAB.

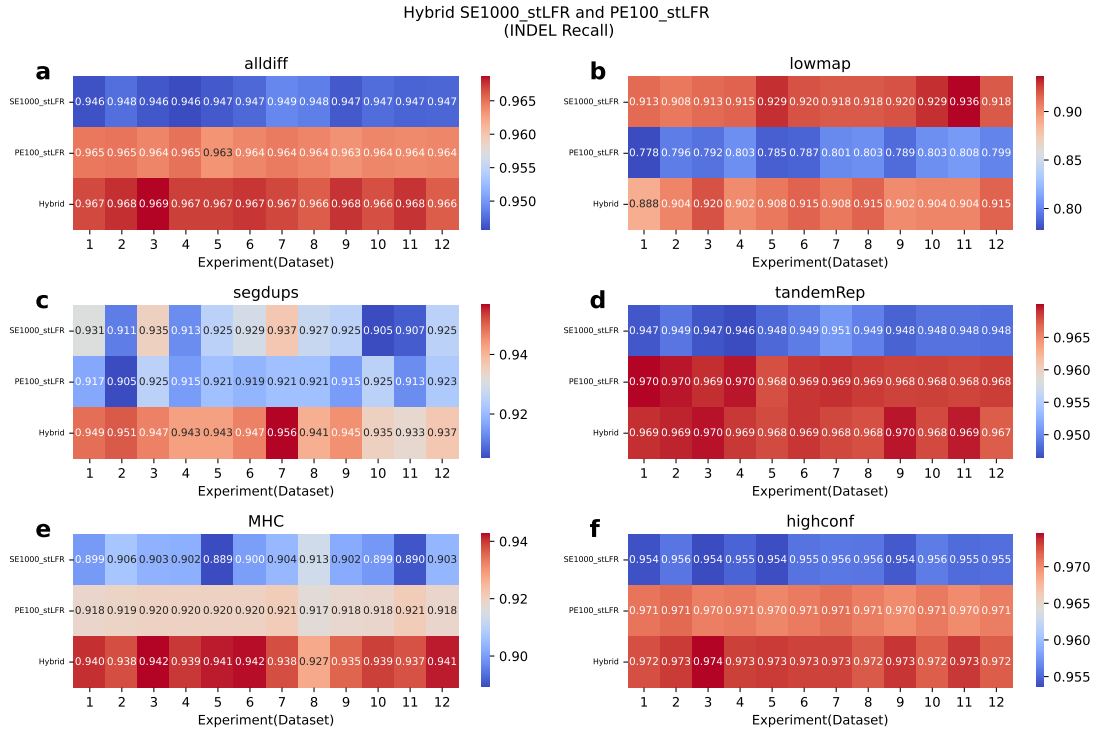


Figure S14: Recall heatmap for INDEL calling on hybrid libraries by combining SE1000_stLFR and PE100_stLFR reads. *hap.py* v0.3.15 was applied to benchmark the INDEL callset against the GIAB ground truth callset. **a.** Recall for all difficult (alldiff) regions. **b.** Recall for low-mappability (lowmap) regions. **c.** Recall for segmentation duplication (segdup) regions. **d.** Recall for tandem and homopolymer repeat (tandemRep) regions. **e.** Recall for MHC regions. **f.** Recall for all high-confidence (highconf) regions defined by GIAB.

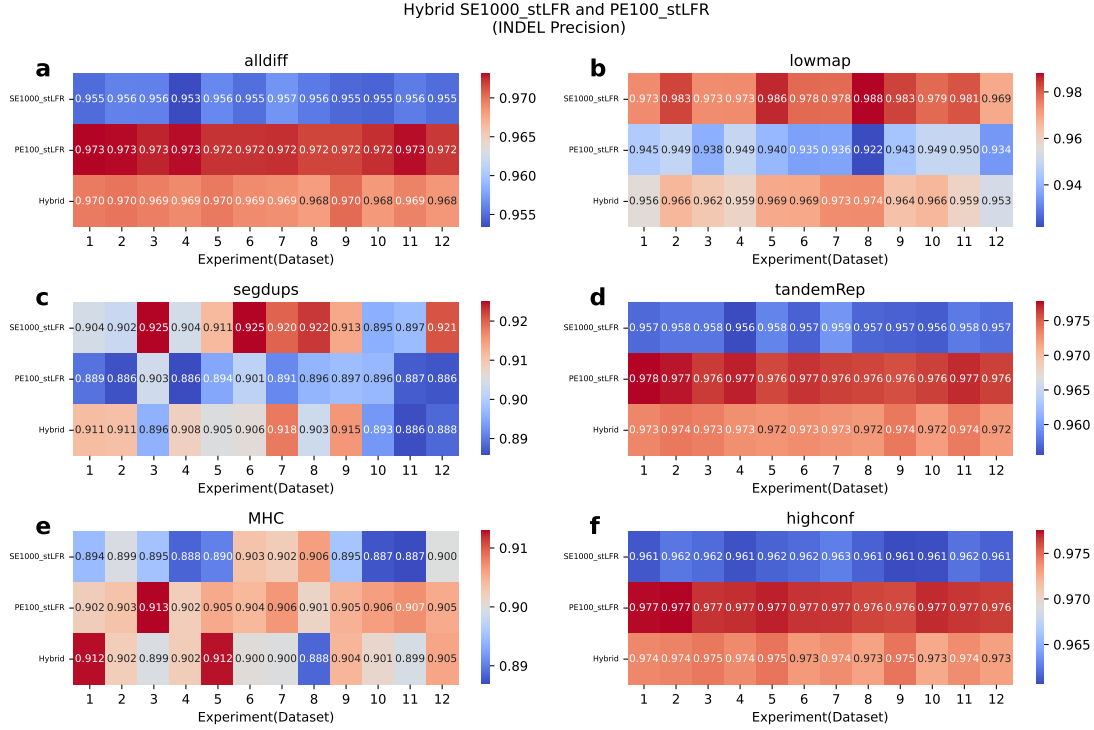


Figure S15: **Precision heatmap for INDEL calling on hybrid libraries by combining SE1000_stLFR and PE100_stLFR reads.** *hap.py* v0.3.15 was applied to benchmark the INDEL callset against the GIAB ground truth callset. **a.** Precision for all difficult (alldiff) regions. **b.** Precision for low-mappability (lowmap) regions. **c.** Precision for segmentation duplication (segdup) regions. **d.** Precision for tandem and homopolymer repeat (tandemRep) regions. **e.** Precision for MHC regions. **f.** Precision for all high-confidence (highconf) regions defined by GIAB.

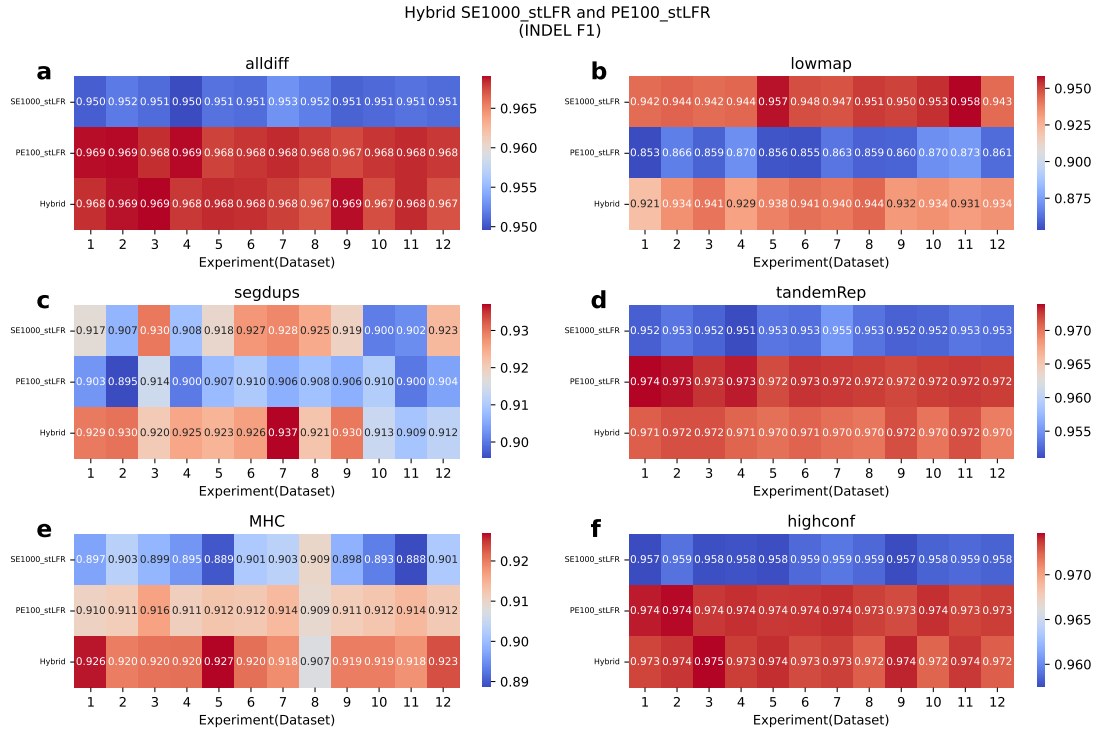


Figure S16: **F1 heatmap for INDEL calling on hybrid libraries by combining SE1000_stLFR and PE100_stLFR reads.** *hap.py* v0.3.15 was applied to benchmark the INDEL callset against the GIAB ground truth callset. **a.** F1 for all difficult (alldiff) regions. **b.** F1 for low-mappability (lowmap) regions. **c.** F1 for segmentation duplication (segdup) regions. **d.** F1 for tandem and homopolymer repeat (tandemRep) regions. **e.** F1 for MHC regions. **f.** F1 for all high-confidence (highconf) regions defined by GIAB.

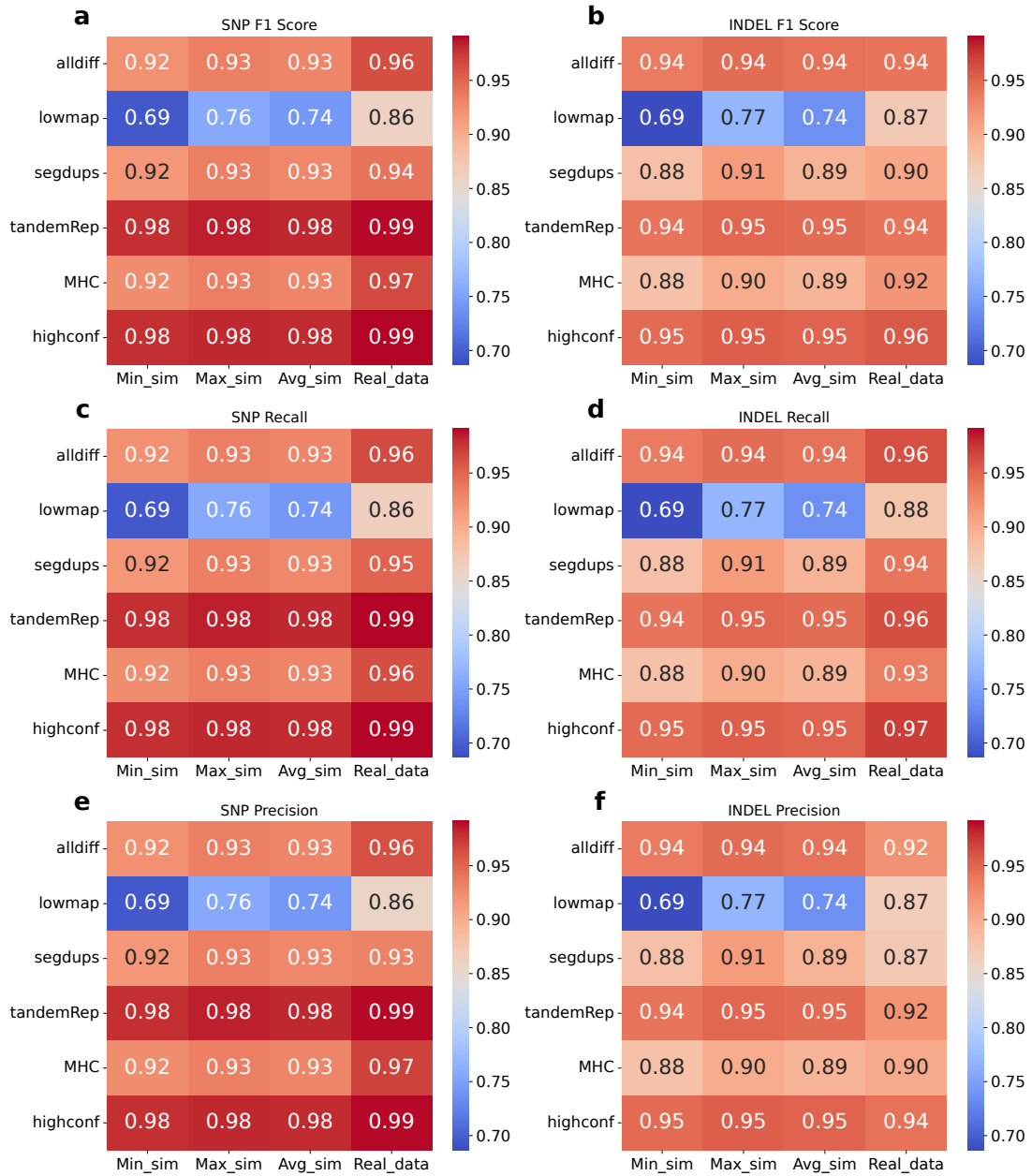


Figure S17: **Simulated and real PE100_stLFR SNP INDEL calling performance comparison** *hap.py* v0.3.15 was applied to benchmark the SNP and INDEL callset against the GIAB ground truth callset. **a.** F1 for SNP calling. **b.** F1 for INDEL calling. **c.** Recall for SNP calling. **d.** Recall for INDEL calling. **e.** Precision for SNP calling. **f.** Precision for INDEL calling

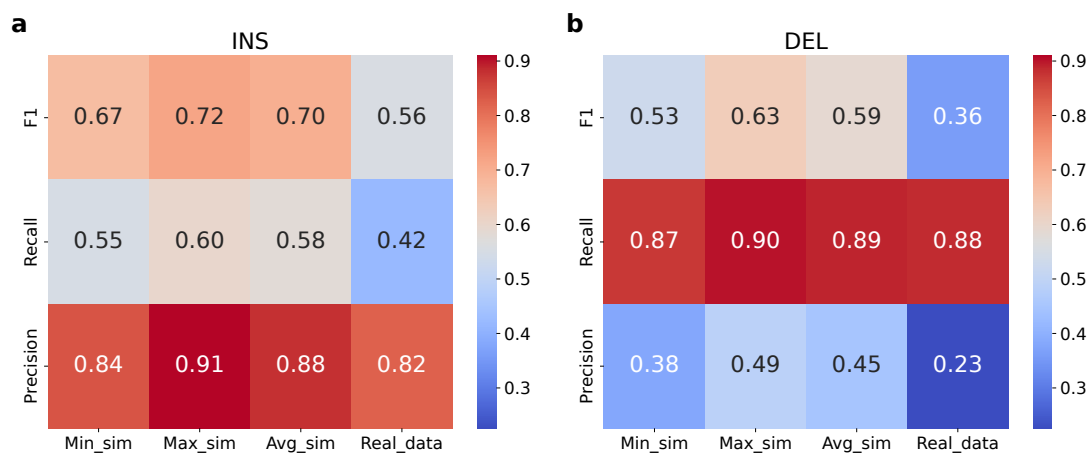


Figure S18: **Simulated and real PE100_stLFR SV calling performance comparison** *Truvari* v4.0.0 was applied to benchmark the SV callset against the GIAB ground truth callset. **a.** F1, recall, precision for insertion SV calling. **b.** F1, recall, precision for deletion SV calling.

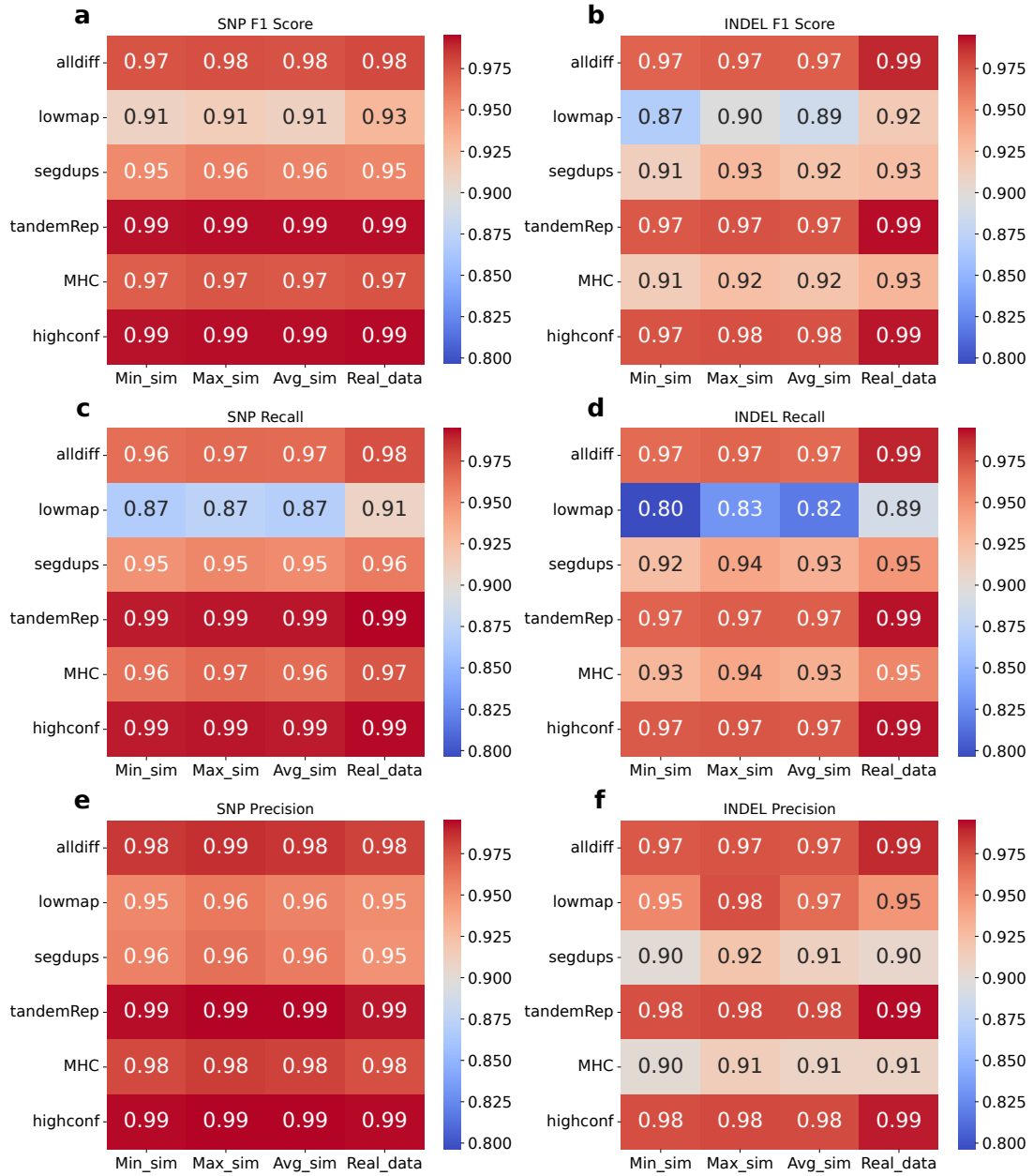


Figure S19: **Simulated and real PE100_stLFR SNP INDEL calling performance comparison** *hap.py* v0.3.15 was applied to benchmark the SNP and INDEL callset against the GIAB ground truth callset. **a.** F1 for SNP calling. **b.** F1 for INDEL calling. **c.** Recall for SNP calling. **d.** Recall for INDEL calling. **e.** Precision for SNP calling. **f.** Precision for INDEL calling