Table of Contents

Supplementary

Supplementary Methods

Comparing structural variants (SV) between different tools

SV call sets from six different tools (pbsv, falcon, smrtsv, sniffles, lumpy, and delly) were downloaded from the GIAB GitHub page (https://github.com/genome-in-a-bottle). For each call set, the distance between SVs were calculated. SVs with any coordinates within 1Kb of each other were considered as overlapping between call sets. Plotting the size of the SVs that “overlapped” showed a high correlation (data not shown). Supplementary Figure S1 shows the percentage of overlap between deletions in the call sets. For example, 86.4% of the deletions identified by MsPAC were found by pbsv, and 84.0% of the deletions identified by pbsv were found by MsPAC. Supplementary Figure 2 shows the number of unique deletions found in each call set. For example, 874 deletions identified by MsPAC were not found by pbsv, and 981 deletions identified by pbsv were not found by MsPAC.

Assessing accuracy of SVs between different tools using Nanopore

Insertion events called by pbsv, falcon, MsPAC and smrtsv were evaluated by comparing against Oxford Nanopore Technology (ONT) reads derived from the AJ sample (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/Ultralong_OxfordNanopore/combined_2018-05-18/combined_2018-05-18.fastq.gz). In each case, the predicted insertion sequence from each caller was inserted *in silico* into the reference at the predicted location, ONT reads were aligned both to the reference and the created insertion reference, and an alignment score was calculated for each. Alignment scores were calculated as the alignment length minus the number of gaps divided by the alignment length. A predicted insertion was considered false if the majority of ONT reads aligned preferentially to the reference. The same procedure was performed for deletions identified by pbsv, falcon, MsPAC, smrtsv, sniffles, lumpy, and delly. Supplementary Table 6 shows the accuracy score for each tool.

Extracting phased intervals

We obtained phased SNVs from 10X (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/10XGenomics_ChromiumGenome_LongRanger2.0_06202016/HG002_NA24385_son/). To extract haplotype intervals we used WhatsHap version 0.18, with the command: whatshap stats <input_vcf> --block-list <regions.bed>”.

Assessing accuracy of phased assembly using Illumina insert libraries

Two different Illumina datasets, 2x250bp (350bp insert) paired-end reads and 6Kb mate pairs reads, were downloaded from ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/NIST_Illumina_2x250bps/ and ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/NIST_Stanford_Illumina_6kb_matepair/fastqs. Each pair of the mate pairs and paired-end reads were aligned to the HG002 assembly separately using BWA mem. The number of mates mapping to the same haplotype were first determined. The mapping was then filtered to select alignments with a mapping quality score greater than 30. The number of mates mapping to the same haplotype increased to 99.6% and 98.5% when the filter was applied to the 2x250bp Illumina library and the 6Kb mate pair Illumina library.

Supplementary Tables

		Example		
Observation	**Explanation**	**Hap 1 base**	**Hap 2 base**	**Reference base**
0	no mutation	A	A	A
1	heterozygous SNV on haplotype 1	A	T	T
2	homozygous SNV	A	A	T
3	heterozygous SNV on haplotype 2	A	T	A
4	heterozygous mulit-allele	A	T	C
5	heterozygous insertion on haplotype 1	A	-	-
6	heterozygous deletion on haplotype 1	-	A	A
7	heterozygous deletion on haplotype 2	A	-	A
8	heterozygous deletion/multi-allele	A	-	T
9	heterozygous insertion on haplotype 2	-	A	-
10	homozygous insertion	A	A	-
11	homozygous insertion/multi-allele	A	T	-
12	homozygous deletion	-	-	A
13	heterozygous deletion/multi-allele	-	A	T
14	Gap sequence	A	A	N

Supplementary Table S1. Definitions of observations in the multiple sequence alignment. Each state models 15 observations. An observation is a column in the multiple sequence alignment between the reference and both haplotypes. The table shows the modelled observations and an example of an observation.

				Current state													
				Insertion						Deletion						Complex	Normal
				Normal			Complex			Normal			Complex				
				1|1	0|1	1|0	1|1	0|1	1|0	1|1	0|1	1|0	1|1	0|1	1|0		
Next state	Insertion	Normal	1|1	.9999													6e-14
			0|1		.9999												6e-14
			1|0			.9999											6e-14
		Complex	1|1				.9999										6e-14
			0|1					.9999									6e-14
			1|0						.9999								6e-14
	Deletion	Normal	1|1							.9999							6e-14
			0|1								.9999						6e-14
			1|0									.9999					6e-14
		Complex	1|1										.9999				6e-14
			0|1											.9999			6e-14
			1|0												.9999		6e-14
	Complex															.9999	6e-14
	Normal			.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	0.99999999999922

Supplementary Table S2. Transition probabilities of Hidden Markov Model (HMM) to call SVs. The table shows the transitional probabilities between the 14 states. Every entry in the table that is not filled are states that are not connected in the HMM. Every state except the normal state can transition to itself or to the normal state, as represented in the table. The normal state can transition into any other state. We chose balanced probabilities for entering events so as to not bias the choice of event types. Transition probability to the normal state was set to .0001 to encourage longer SVs.

	State													
	Insertion						Deletion						Com plex	Nor- mal
	Normal			Complex			Normal			Complex				
Obs.	1|1	0|1	1|0	1|1	0|1	1|0	1|1	0|1	1|0	1|1	0|1	1|0		
0	.004	.004	.004	**.01**	**.01**	**.01**	.004	.004	.004	**.01**	**.01**	**.01**	.006	**.95**
1	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.006	.004
2	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.006	.004
3	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.006	.004
4	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.006	.004
5	.004	**.95**	.004	.004	**.94**	.004	.004	.004	.004	.004	.004	.004	**.158**	.004
6	.004	.004	.004	.004	.004	.004	.004	.004	**.95**	.004	.004	**.94**	**.158**	.004
7	.004	.004	.004	.004	.004	.004	.004	**.95**	.004	.004	**.94**	.004	**.158**	.004
8	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.006	.004
9	.004	.004	**.95**	.004	.004	**.94**	.004	.004	.004	.004	.004	.004	**.158**	.004
10	**.95**	.004	.004	**.94**	.004	.004	.004	.004	.004	.004	.004	.004	**.158**	.004
11	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.006	.004
12	.004	.004	.004	.004	.004	.004	**.95**	.004	.004	**.94**	.004	.004	**.158**	.004
13	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.006	.004
14	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.006	.004

Supplementary Table S3. Matrix with the observation probabilities. The emission probability (observation probability) for each observation and for each state is listed in the table. In this study, for non-complex events we set the probability of observing the expected state to 95% (lowering it to 94% for complex variants and allowing for more “normal” bases). For fully “complex” events, we distributed the various insertion/deletion states evenly under the assumption no specific event-type was preferred.

Haplotype	Coverage (no ambig. reads)	Coverage (ambig. reads included)	Bases assembled	% of chr1-22 assembled	N50	Longest contig
1	26.6	52.4	2.45GB	91.4%	4.3MB	22.7MB
2	26.8	52.5	2.47GB	91.8%	4.2MB	21.3MB

Supplementary Table S4. MsPAC haplotype assembly statistics. The table shows the assembly statistics for each haplotype assembled by MsPAC.

SV	Homozygous	Heterozygous (SV in haplotype 1)	Heterozygous (SV in haplotype 2)	Validation rate using ONT	SVs containing TRs*
Deletion	2,375	2,021	2,014	95% (2,947/3,113)	2,780
Insertion	4,420	2,367	2,465	87% (3,438/3,930)	5,431
Complex Deletion	421	729	337	72% (150/207)	327
Complex Insertion	646	721	288	74% (147/198)	653
Complex	112	NA	NA	NA	54

Supplementary Table S5. SV statistics produced by MsPAC. The table shows the number of different SVs identified separated by genotype, with the validation rate using ONT reads.

MsPAC step	Max runtime per job	Max resources per job	Number of jobs	Total CPU runtime
Phasing	< 4 hours	2 GBs, 1 core	22 (or # of chromosomes)	~ 30 hours
Separating raw PacBio reads	< 20 hours	40GBs, 1 core	22 (or # of chromosomes)	~100 hours
Assembly	< 30 min	8 GBs, 1 core	< 10,000	~2,260 hours
SV detection	< 20 min	2 GBs. 1 core	~ 10,000	~350 hours

Supplementary Table S6. Runtime and CPU resources for different MsPAC steps for the HG002 dataset. The table shows the runtime and CPU resources needed for each job for each specific MsPAC step and the number of jobs per step. The whole process takes 3 – 8 days depending on the number of jobs distributed to the cluster and availability of the cluster.

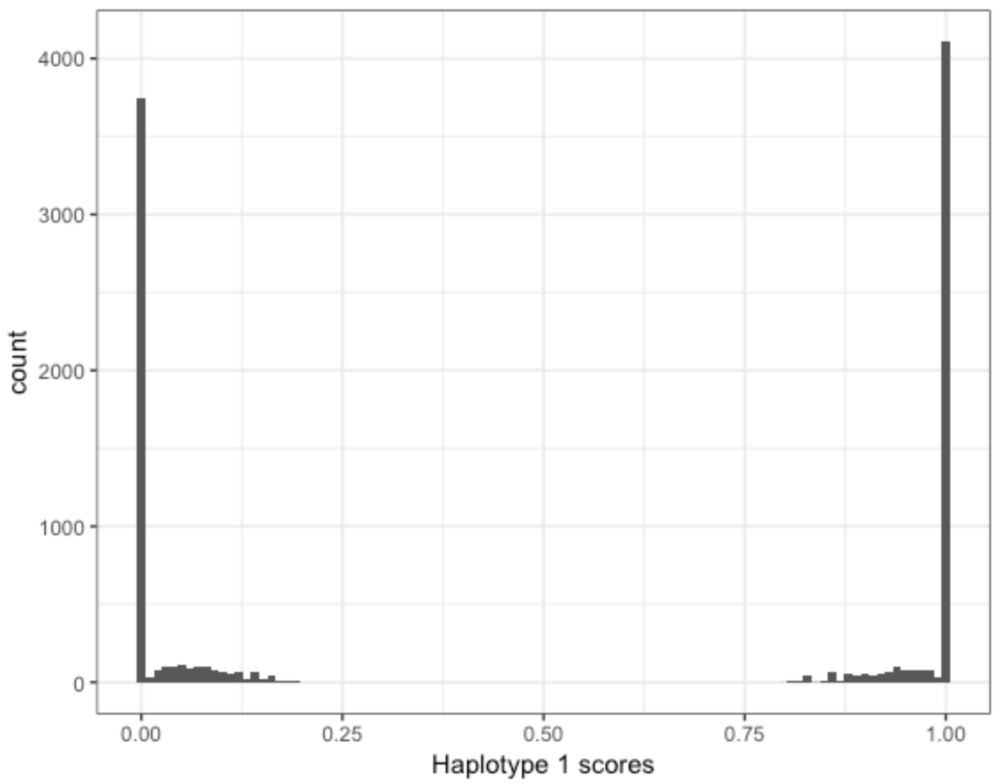
Call set (version, if found)	URL
PbHoney (svn revision 107)	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/BCM_PacBio_PBHoney_Sep.8.2016_Filt/Calls.bed
smrtsv (June 2016)	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/Chaisson_PacBio_smrt-sv.dip_Jun2016/deletions.bed, ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/Chaisson_PacBio_smrt-sv.dip_Jun2016/insertions.bed
Pbsv (v0.1-prerelease)	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/PacBio_pbsv_05052017/hg19.HG002.pbsv.vcf.gz
sniffles	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/Baylor_sniffles_05092017/all_reads.fa.giab_h002_ngmlr-0.2.3_mapped.bam.sniffles1kb_auto_noalts.vcf.gz
delly	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/DNAnexus_AndrewC_Illumina_Callers_Sep2016/HG002/HG002.140528_D00360_0018_AH8VC6ADXX.realigned.recalibrated.delly.deletion.vcf.gz
lumpy	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/DNAnexus_AndrewC_Illumina_Callers_Sep2016/HG002/HG002.140528_D00360_0018_AH8VC6ADXX.realigned.recalibrated.lumpy.vcf
GIAB Tier 1 call set	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6/HG002_SVs_Tier1_v0.6.vcf.gz
Falcon	ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/MtSinai_PacBio_Assembly_falcon_03282016/

Supplementary Table S7. Call sets used for comparison. The table shows the URL of different SVs call sets. The Falcon call set was generated by applying the SV-calling pipeline from MsPAC on the Falcon assembled contigs.

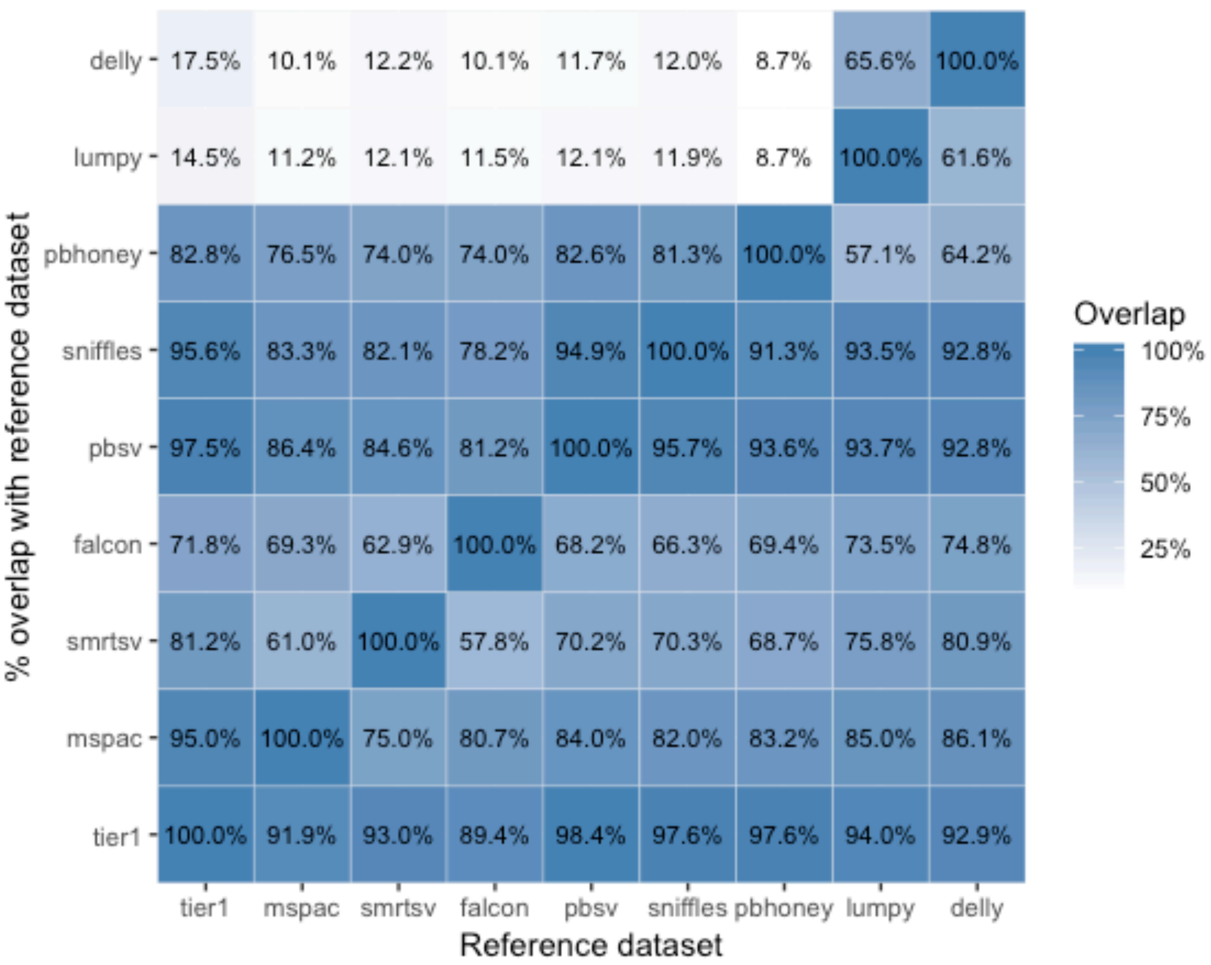
SV	Tool	Incorrect	Total	Precision
Deletion	Falcon	149	2474	94.0%
	MsPAC	166	3113	94.7%
	smrtsv	138	2850	95.2%
	Pbhoney	**207**	2925	92.9%
	pbsv	78	**3345**	97.6%
	sniffles	76	3115	97.6%
	delly	8	444	**98.2%**
	lumpy	15	412	96.4%
Insertion	Falcon	442	3432	87.1%
	MsPAC	492	3930	87.5%
	smrtsv	405	3562	**88.6%**
	pbsv	**670**	**4531**	85.2%

Supplementary Table S8. Accuracy of SVs for different tools using ONT reads. The table shows the insertion and deletion accuracy for each of the tools tested

Supplementary Figures

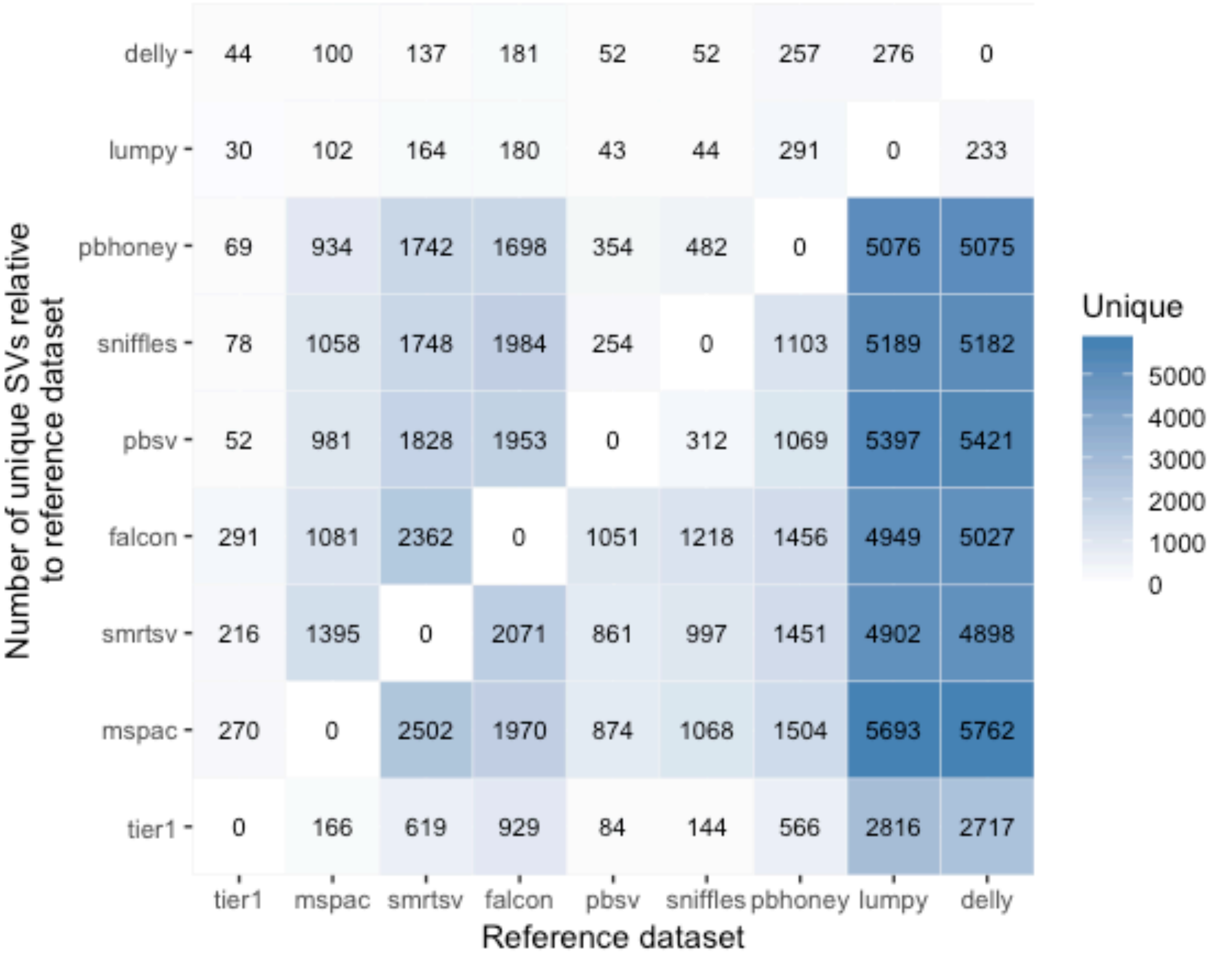


Supplementary Figure S1. Histogram of haplotype 1 scores across all reads. For each read, we plot $\frac{P(r|h_r=1)}{P(r|h_r=1)+P(r|h_r=2)}$, which shows a bimodal distribution Reads with a score near 1 are likely to be from haplotype 1 and those with a value near 0 are likely to be derived from haplotype 2.



	tier1	mspac	smrtsv	falcon	pbsv	sniffles	pbhoney	lumpy	delly
delly	17.5%	10.1%	12.2%	10.1%	11.7%	12.0%	8.7%	65.6%	100.0%
lumpy	14.5%	11.2%	12.1%	11.5%	12.1%	11.9%	8.7%	100.0%	61.6%
pbhoney	82.8%	76.5%	74.0%	74.0%	82.6%	81.3%	100.0%	57.1%	64.2%
sniffles	95.6%	83.3%	82.1%	78.2%	94.9%	100.0%	91.3%	93.5%	92.8%
pbsv	97.5%	86.4%	84.6%	81.2%	100.0%	95.7%	93.6%	93.7%	92.8%
falcon	71.8%	69.3%	62.9%	100.0%	68.2%	66.3%	69.4%	73.5%	74.8%
smrtsv	81.2%	61.0%	100.0%	57.8%	70.2%	70.3%	68.7%	75.8%	80.9%
mspac	95.0%	100.0%	75.0%	80.7%	84.0%	82.0%	83.2%	85.0%	86.1%
tier1	100.0%	91.9%	93.0%	89.4%	98.4%	97.6%	97.6%	94.0%	92.9%

Supplementary Figure S2. Overlap between deletions made using different SV callers and GIAB Tier 1 deletion call set. The matrix shows the overlap of deletions between different tools and GIAB Tier 1 deletions, with each cell showing the percent pairwise overlap.



	tier1	mspac	smrtsv	falcon	pbsv	sniffles	pbhoney	lumpy	delly
delly	44	100	137	181	52	52	257	276	0
lumpy	30	102	164	180	43	44	291	0	233
pbhoney	69	934	1742	1698	354	482	0	5076	5075
sniffles	78	1058	1748	1984	254	0	1103	5189	5182
pbsv	52	981	1828	1953	0	312	1069	5397	5421
falcon	291	1081	2362	0	1051	1218	1456	4949	5027
smrtsv	216	1395	0	2071	861	997	1451	4902	4898
mspac	270	0	2502	1970	874	1068	1504	5693	5762
tier1	0	166	619	929	84	144	566	2816	2717

Supplementary Figure S3. Number of unique deletions between call sets. The matrix shows the unique deletions identified by each tool and GIAB Tier 1 deletion call set on the Y axis when compared the reference dataset in the X axis.