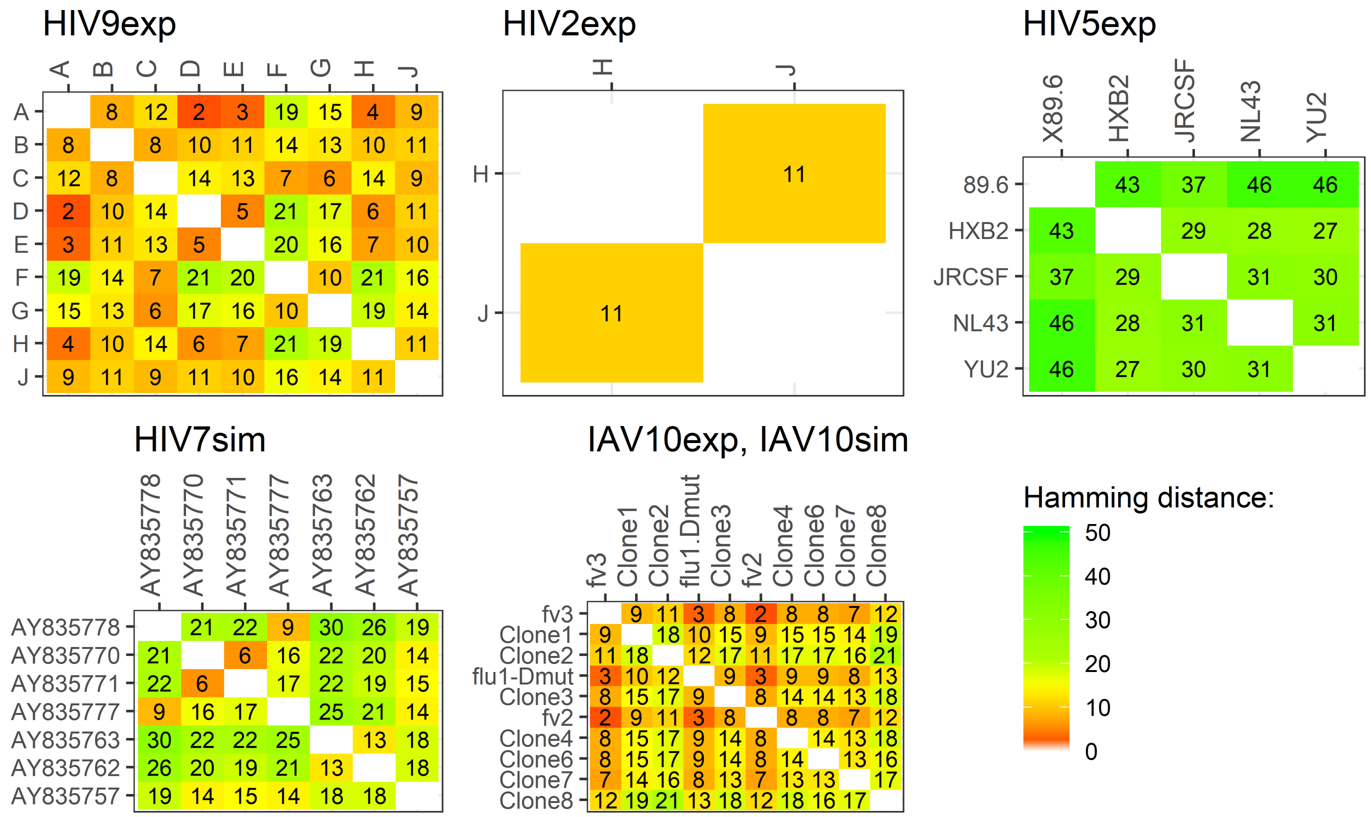
Supplementary Material

Supplementary Results



HIV9exp

	A	B	C	D	E	F	G	H	J
A		8	12	2	3	19	15	4	9
B	8		8	10	11	14	13	10	11
C	12	8		14	13	7	6	14	9
D	2	10	14		5	21	17	6	11
E	3	11	13	5		20	16	7	10
F	19	14	7	21	20		10	21	16
G	15	13	6	17	16	10		19	14
H	4	10	14	6	7	21	19		11
J	9	11	9	11	10	16	14	11	

HIV2exp

	H	J
H		11
J	11	

HIV5exp

	X89.6	HXB2	JRCSF	NL43	YU2
89.6		43	37	46	46
HXB2	43		29	28	27
JRCSF	37	29		31	30
NL43	46	28	31		31
YU2	46	27	30	31	

HIV7sim

	AY835778	AY835770	AY835771	AY835777	AY835763	AY835762	AY835757
AY835778		21	22	9	30	26	19
AY835770	21		6	16	22	20	14
AY835771	22	6		17	22	19	15
AY835777	9	16	17		25	21	14
AY835763	30	22	22	25		13	18
AY835762	26	20	19	21	13		18
AY835757	19	14	15	14	18	18	

IAV10exp, IAV10sim

	fv3	Clone1	Clone2	flu1.Dmut	Clone3	fv2	Clone4	Clone6	Clone7	Clone8
fv3		9	11	3	8	2	8	8	7	12
Clone1	9		18	10	15	9	15	15	14	19
Clone2	11	18		12	17	11	17	17	16	21
flu1-Dmut	3	10	12		9	3	9	9	8	13
Clone3	8	15	17	9		8	14	14	13	18
fv2	2	9	11	3	8		8	8	7	12
Clone4	8	15	17	9	14	8		14	13	18
Clone6	8	15	17	9	14	8	14		13	16
Clone7	7	14	16	8	13	7	13	13		17
Clone8	12	19	21	13	18	12	18	16	17	

Figure S1. Pairwise Hamming distances between variants in the experimental (exp) and simulated (sim) datasets HIV9exp, HIV2exp, HIV5exp, HIV7sim, IAV10sim, and IAV10exp.

PacBio # of Reads	Method	Variant	fv3	Clone1	Clone2	flu1-Dmut	Clone3	fv2	Clone4	Clone5	Clone6	Clone7	FP
		True Freq.,%	50	25	12.5	6.25	3.125	1.56	0.78	0.39	0.19	0.097	
33.5K (all)	CliqueSNV	Match	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	0
		Freq., %	52.6	23.7	12.6	6.4	2.3	1.17	0.7	0.35	0.12	0.051	0
	2SNV	Match	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	1
		Freq., %	51.8	23.7	12.5	6.4	2.3	1.2	0.7	0.3	0.1	0	1.0
	PredictHaplo	Match	✓	✓	✓	×	✓	×	✓	✓	×	×	0
		Freq.,%	56.7	23.8	13.7	0	3.1	0	1.5	1.2	0	0	0
Subsampling													
16K	CliqueSNV	Match,%	100	100	100	100	100	90	100	100	100	20	0.1
		Freq., %	52.9	23.7	12.5	6.4	2.3	1.19	0.71	0.32	0.12	0.69	1.15
	2SNV	Match,%	100	100	100	100	100	100	100	100	0	0	0.2
		Freq., %	52.4	23.7	12.5	6.4	2.3	1.1	0.7	0.3	0	0	0.6
	PredictHaplo	Match	100	100	100	70	100	0	100	40	0	0	0.3
		Freq.,%	54.2	23.5	13.1	6.0	2.9	0	1.4	1.0	0	0	0.5
8K	CliqueSNV	Match,%	100	100	100	100	100	90	100	100	30	0	0
		Freq., %	52.8	23.6	12.5	6.5	2.3	1.2	0.7	0.35	0.16	0	0
	2SNV	Match,%	100	100	100	100	100	100	100	0	0	0	0
		Freq., %	53.1	23.7	12.5	6.5	2.3	1.25	0.7	0	0	0	0
	PredictHaplo	Match,%	100	100	100	0	100	0	100	20	0	0	0.2
		Freq.,%	58.1	24.0	12.7	0	3.1	0	1.6	1.3	0	0	0.5
4K	CliqueSNV	Match,%	100	100	100	100	100	80	100	40	0	0	0
		Freq., %	53.3	23.7	12.3	6.4	2.4	1.19	0.7	0.39	0	0	0
	2SNV	Match,%	100	100	100	100	100	100	20	0	0	0	0
		Freq., %	53.7	23.7	12.3	6.5	2.4	1.2	0.9	0	0	0	0
	PredictHaplo	Match,%	100	100	100	0	70	0	10	0	0	0	0.3
		Freq.,%	60.1	23.9	12.8	0	3.5	0	2.5	0	0	0	0.5

Table S1. Comparison of CliqueSNV, 2SNV and PredictHaplo on full and sub-sampled data ***(PacBio, experimental)***. For all 33.5K reads, the sign "✓" (respectively, "×") denotes fully matched (respectively, unmatched) true variant and the column FP reports the number of incorrectly predicted variants (false positives) and their total frequency. For each sub-sample size (16K,. . . ,4K), the table reports the percent of runs when a variant is completely matched and its average frequency over runs when the variant was detected. Similarly, the column FP reports the average number of false positive variants and their average total frequency. Colors indicate the percent of matched variants: green - high percent, red - low percent.

Benchmark	Length	Consensus	CliqueSNV			PredictHaplo	aBayesQR
			2%	5%	10%		
HCV10sim	1K	13.52	64.12	72.59	65.86	314.87	did not finish
	2K	13.85	169.16	133.06	108.46	972.41	did not finish
	5K	16.79	3666.76	3117.49	221.70	6472.83	did not finish
	full-length	15.27	3703.01	3559.10	483.77	58509.17	did not finish
ZIKV3sim	1K	34.61	81.88	91.21	91.50	88.76	4409.53
	2K	31.57	104.71	115.81	106.82	342.31	did not finish
	5K	33.62	161.90	156.31	160.64	1775.49	did not finish
	full-length	35.20	271.55	281.47	284.54	12114.49	did not finish
ZIKV15sim	1K	13.33	114.42	117.75	139.08	314.87	did not finish
	2K	13.16	148.40	153.95	147.76	342.31	did not finish
	5K	13.70	337.82	229.16	166.85	1775.49	did not finish
	full-length	13.66	10305.01	604.60	286.19	12114.49	did not finish
HIV5full	1K	21.60	247.84	215.70	208.73	155.11	24462.81
	2K	20.18	1282.03	460.03	374.76	459.40	28820.99
	5K	19.77	5291.37	1787.24	337.52	2982.96	did not finish
	full-length	20.26	8084.50	4970.50	1153.09	14404.43	did not finish
Average over all benchmarks		20.63	2127.16	1004.12	271.08	7071.21	21628.58

Table S2. Running time of performed experiments (seconds) for full-length benchmarks.

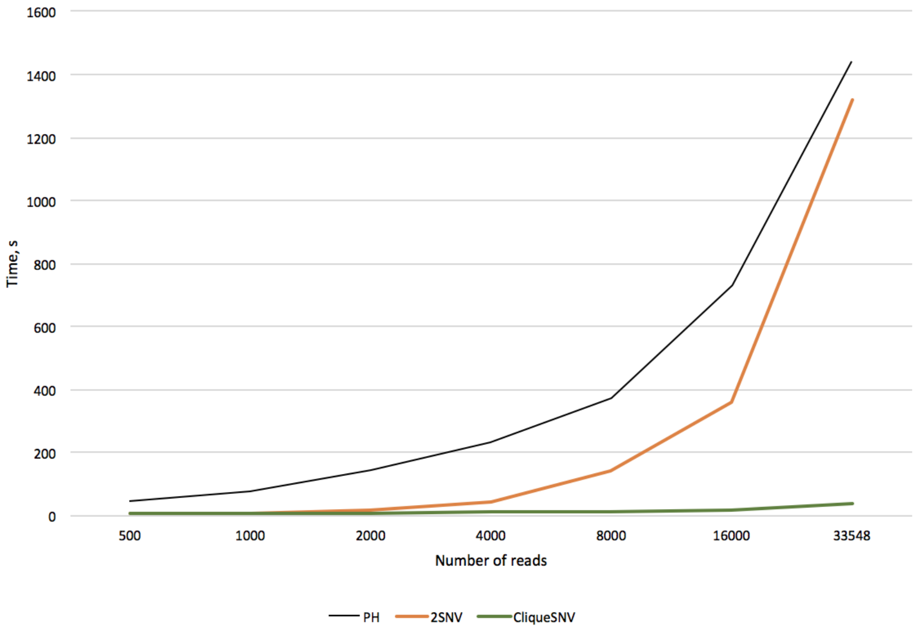


Figure S2. Runtimes of PredictHaplo (PH), 2SNV and CliqueSNV on datasets with different read sizes.

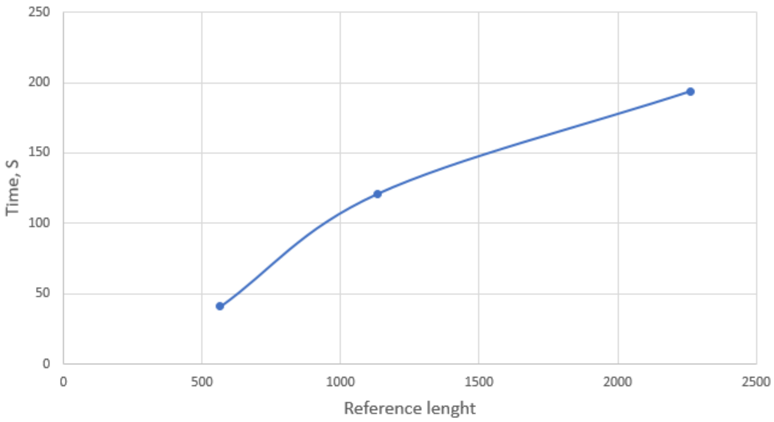


Figure S3. CliqueSNV runtime on datasets with different reference length and same coverage (about 1M reads in total).

Supplementary Methods

Benchmark preparation

We used 50,000 total copies of plasmid DNA from these nine constructs as input for a nested PCR reaction to amplify the polymerase region using the following primary and nested primers respectively: HIV-B PRO-OUT.3F, 5' CCT CAG ATC ACT CTT TGG CAA CG 3' and HIV-RT 215/219.3R, 5' CTT CTG TAT GTC ATT GAC AGT CC 3' Nested PCR: HIV-B PR/RT.2F, 5' CTT TGG CAA CGA CCC CTY GTC CA 3' and HIV-RT 181-190.1.4R, 5' ATC AGG ATG GAG TTC ATA ACC CA 3'.

The primary and nested PCRs were done using 94°C for four minutes, followed by 40 cycles of 94°C for one minute, 50°C for 30 seconds, and 72°C for two minutes and a final extension at 72°C for five minutes.

We created two plasmid mixtures to generate artificial mixtures simulating clinical specimens containing many variants at different virus levels. The mixtures comprised nine and two plasmids with varying copy numbers of each plasmid.

PCR reactions were generated and purified using the QIAquick PCR purification kit. (Qiagen, Valencia CA) The purified amplicons (10 ng) were subsequently used for NGS library construction using the Nextera XT DNA Library Prep kit (Illumina Inc., San Diego, CA). Libraries were pooled, and enriched for 900-1,000-bp fragments using magnetic bead based size selection (AMPure XP, Beckman Coulter, Brea, CA) and sequenced on a MiSeq v3 (600-cycle) flow cell on the Miseq system (Illumina Inc., San Diego CA).

Pseudocode of the CliqueSNV algorithm

Algorithm 1 CliqueSNV Algorithm

```
Step 1: finding linked and forbidden SNV pairs
    Split the read alignment $M_{L\times N}$ into binary matrix $4M$
    Construct a compact representation of the binary matrix $4M$
    For each $I, J \in \{1,\dots,4L\}$ find $O^{IJ}$ and $O^{IJ}_{22}$, where
        $O^{IJ}$ = # of reads covering both $I$ and $J$
        $O^{IJ}_{22}$ = # of reads with both minor SNVs
        If $O^{IJ}_{22} > \varepsilon O^{IJ}$ compute $p$-value (default $\varepsilon = 0.0003$)
    Find all linked SNV pairs with the adjusted p-value $< 1\%$
Step 2: constructing the SNV graph
    Filter out 10% of the most erroneous PacBio reads
    Construct the SNV graph $G = (V, E)$, where
        $V = \{1,\dots,4L\}$, and $E$ are links between minor SNVs
Step 3: finding maximal cliques in the SNV graph using Bron-Kerbosch algorithm
Step 4: merging cliques in the clique graph with forbidden pairs
    Find the clique graph $C_G$ with pairs.
    Find all maximal connected subgraphs in $C_G$.
    Merge all cliques inside each maximal connected subgraph.
Step 5: partitioning reads between merged cliques and finding consensus haplotypes
    Find the set $S$ of all positions that belong to at least one clique.
    Make an empty clique on $S$.
    Assign each read to the closest clique.
    Find the consensus $v(q)$ of all assigned reads for each $q$.
Step 6: estimating haplotype frequencies by expectation-maximization algorithm
```

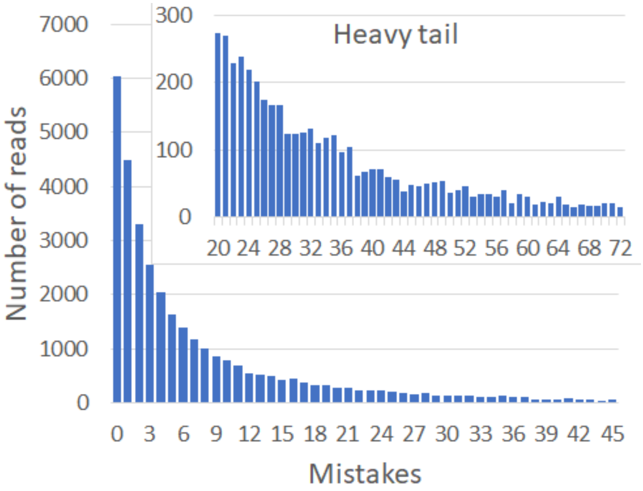


Figure S4. A typical distribution of errors in PacBio reads. The heavy tail indicates that a significant portion of errors is accumulated by a relatively small number of reads.

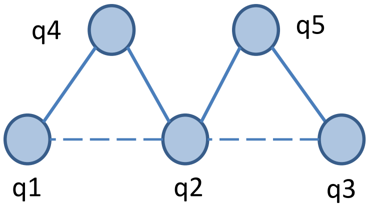


Figure S5. The clique graph C_G with 5 vertice corresponding to cliques in G, 4 edges and two forbidden pairs (q_1,q_2) and (q_2,q_3). There 3 maximal connected subgraphs avoiding forbidden pairs: $\{q_1,q_4\}$ $\{q_4,q_2,q_5\}$ $\{q_5,q_3\}$

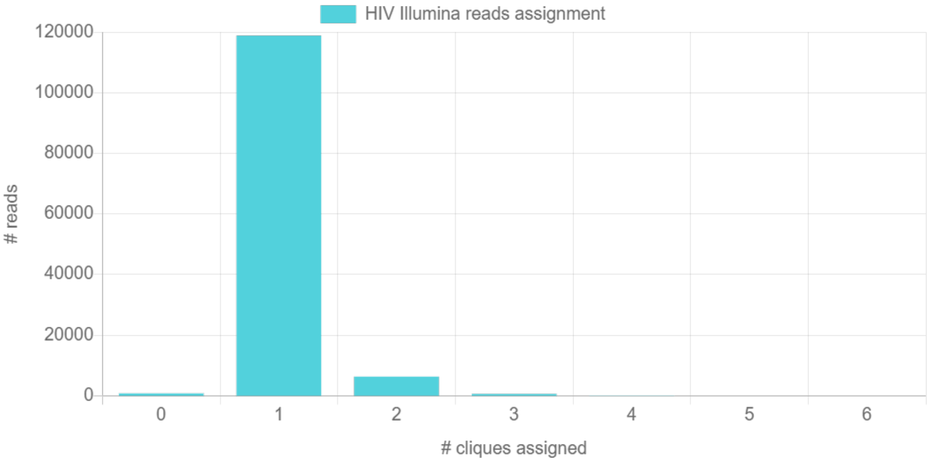


Figure S6. The number of reads assigned to different number of cliques in HIV Illumina dataset.

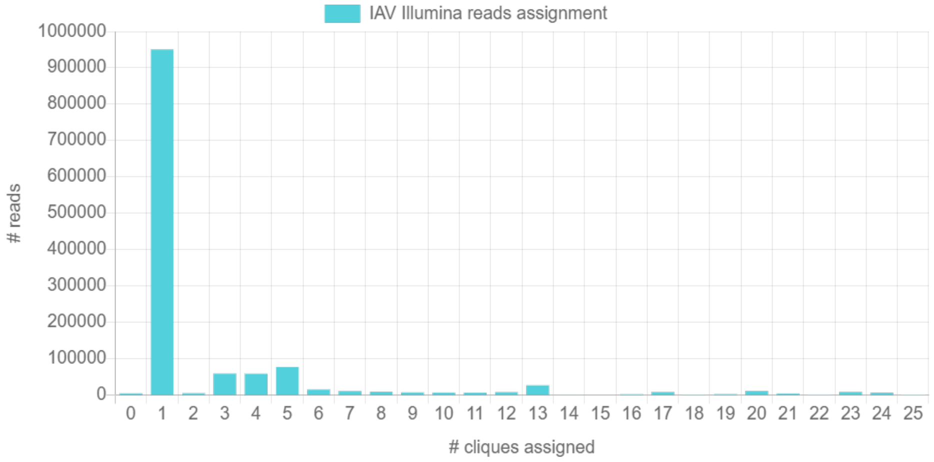


Figure S7. The number of reads assigned to different number of cliques in IAV Illumina dataset.

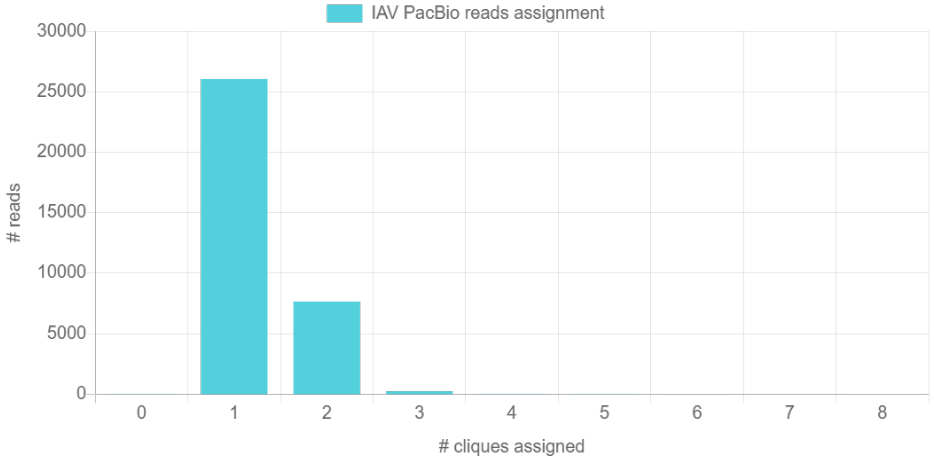


Figure S8. The number of reads assigned to different number of cliques in IAV PacBio dataset.