Table 1. Characteristics of this study cohort.

Patient ID	Gender	HIV	BLx [a]	BLx Background	Clinical Gt[d]
HCVHIV02	Male	+	UCFC [b]	Hemophilia	1a + 1b
HCVHIV03	Male	+	UCFC	Hemophilia	1b
HCVHIV04	Male	+	UCFC	Hemophilia	1b
HCVHIV05	Male	+	UCFC	Hemophilia	1b
HCVHIV06	Male	+	UCFC	Hemophilia	2a
HCVHIV07	Male	+	UCFC	Hemophilia	1b
HCVHIV10	Male	+	UCFC	Hemophilia	Untyped
HCVHIV11	Male	+	UCFC	Hemophilia	1
HCVHIV15	Male	+	UCFC	Hemophilia	2b
HCVHIV17	Male	+	UCFC	Hemophilia	1b
HCVmono15	Male	–	BT [c]	Not Available	1
HCVmono17	Male	–	BT	Traffic Accident	1
HCVmono19	Male	–	–	Unknown, BT (–)	1
HCVmono20	Male	–	BT	Burn Injury	1
HCVmono23	Male	–	–	Unknown, BT (–)	1
HCVmono25	Male	–	–	Unknown, BT (–)	1
HCVmono27	Female	–	–	Needlestick Injury	1
HCVmono28	Female	–	BT	Traffic Accident	1
HCVmono29	Female	–	–	Unknown, BT (–)	1
HCVmono34	Female	–	BT	Caesarean section	1

[a] BLx: Any exposure to blood/blood-related product

[b] UCFC: Unheated coagulation factor concentrates

[c] BT: Whole-blood transfusion

[d] Clinical Gt: Results of clinical genotyping / serotyping

doi:10.1371/journal.pone.0119145.t001

initial denaturation at 94°C for 2 min, followed by 40 cycles of 94°C for 20 sec, 50°C for 30 sec, and 68°C for 4 min. After the 10th cycle, the elongation step was extended in increments of 3 sec per cycle. The final elongation was at 68°C for 20 min. The second round PCR conditions were the same as those of the first round PCR, with the exception of annealing temperature (55°C instead of 50°C). The primers used in this study are listed in S1 Table. An amplicon of around 4.2 kbp was excised from agarose gel and purified using MinElute Gel Extraction Kit (Qiagen), eluted in 20 μl of DNase-free water, and preserved at −20°C for downstream applications.

Illumina MiSeq next-generation sequencing of partial core to NS3 protease region

PCR amplicons were quantified using a Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA) and a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). Paired-end libraries were prepared from 200 ng of DNA using a TruSeq Nano DNA Sample Prep Kit (Illumina, San Diego, CA, USA). Manufacturer's instructions were strictly followed. Size selection using SPRI beads resulted in DNA ligated with adapters at a size distribution of around 800 bp. Eight cycles of PCR were carried out using barcoded primers, thereby the DNA insert fragment flanked with adapter sequences was enriched. Purified PCR products were pooled so as to contain an equimolar concentration of each library, and 2 x 300 bp paired-end sequencing was carried out using MiSeq and MiSeq Reagent Kits V3 (Illumina).

Control RNA preparation for estimating MiSeq error rate

Control HCV RNA was prepared by *in vitro* transcription with T3 RNA polymerase (Promega, Madison, WI, USA) and rNTPs (Ambion, Austin, TX, USA) from a linearized plasmid. As a template, a bacterially amplified plasmid (pBSK HC-J1), containing a T3 promoter and a full-length HC-J1 isolate (subtype 1b) sequence (GenBank D10749), kindly gifted by Dr. Tetsuro Suzuki, was used. The prepared RNA was pretreated with TURBO DNase (Invitrogen, Carlsbad, CA, USA), purified using a QIAamp Viral RNA Mini Kit (Qiagen), quantified by NanoDrop spectrophotometer (Thermo Scientific, IL, USA), split into aliquots, and stored at −80°C. Complete digestion of the template plasmid was confirmed by nested RT-PCR omitting the RTase. NGS libraries were prepared in duplicates and sequenced.

MiSeq data accessibility

Illumina MiSeq sequence datasets (in fastq format) are accessible in the DDBJ Sequence Read Archive (http://trace.ddbj.nig.ac.jp/dra/index_e.html) under the Accession Number of DRA002750.

Bioinformatics

All sequence analyses were performed using Geneious 7.1 software (Biomatters Ltd., http://www.geneious.com/), sequence analysis suite implemented in Java. All simulations and custom bioinformatics analyses were carried out using Mathematica version 10.0 (Wolfram Research, Inc., http://www.wolfram.com/mathematica/) unless otherwise specified. R version 3.1 [27] (http://www.r-project.org/) and additional Bioconductor libraries [28] (http://www.bioconductor.org/) were also utilized. All scripts are available upon request. An analysis flowchart was shown in Fig. 1.

MiSeq read quality control and mapping

Generated reads were adaptor-trimmed using the *cutadapt* program [29] (https://code.google.com/p/cutadapt/). Low-quality reads were removed with a threshold of average quality score < 20. Contaminating PhiX control reads were then removed by mapping to the PhiX reference sequence with the BWA program [30] (http://bio-bwa.sourceforge.net/) with parameters of [-l 20-k 5-n 150]. Unmapped reads were binned using barcode sequences and used for downstream analyses. Mapping to the HCV H77 reference genome (GenBank AF01175) was conducted using Geneious default mapper with the parameter setting of "highest-sensitivity". Mapping was iterated three times.

Estimation of pairwise SNV-to-SNV nucleotide distance distribution

Since quasispecies reconstruction (QSR) requires quasispecies-to-quasispecies nucleotide mismatches, or SNVs, as "scaffolds" for concatenating NGS reads that are likely derived from the same quasispecies [20], the distribution of the nucleotide distance between SNVs in any region of interest would be critically important for reliable reconstruction, that is, avoiding the artificial generation of *in silico* recombinants.

Genotype reference sequences were obtained from the HCV Sequence Alignments web tool in The Los Alamos hepatitis C sequence database [31] (http://hcv.lanl.gov/content/sequence/NEWALIGN/align.html) (Alignment type = "Genotype reference", and Year = "2012"). Obtained sequences were pairwise-aligned using MAFFT [32]. Pairwise SNV-to-SNV nucleotide distance distribution was defined as a set of nucleotide distances between mismatched bases in

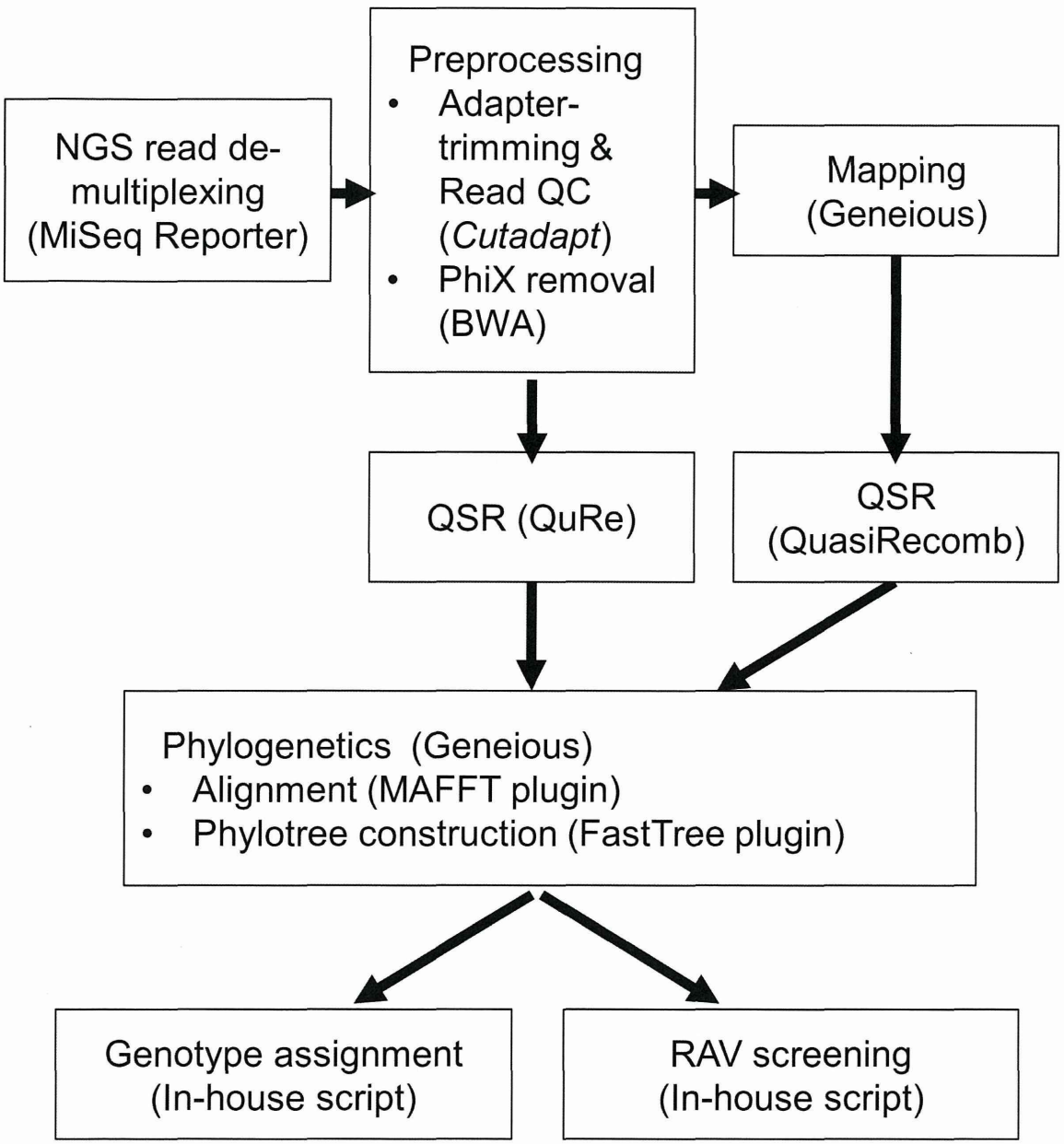

Fig 1. A flowchart of bioinformatics employed in this study.

doi:10.1371/journal.pone.0119145.g001

each pairwise alignment. Pairwise alignments of two reference sequences of the same genotypes were categorized as 'intragenotype' and those of the same subtypes as 'intrasubtype'.

Quasispecies reconstruction

To simultaneously infer geno/subtype and linked amino acid variants, a conventional SNV calling approach is unsatisfactory. Therefore, an alternative approach, quasispecies reconstruction (QSR), was employed in this study. QSR was performed using both QuRe v0.99971 [25]

(http://sourceforge.net/projects/qure/) and QuasiRecomb 1.2 [26] (http://www.cbg.ethz.ch/software/quasirecomb). Since Prosperi et al. [22] previously studied the performance of QuRe, ShoRAH [33] and PredictHaplo (http://cs-wwwarchiv.cs.unibas.ch/personen/roth_volker/HivHaploTyper/index.html), demonstrating QuRe as having a low false positive rate and a reasonably high recall rate among programs compared, ShoRAH and PredictHaplo were omitted from this study. Another set of programs, V-Phaser [34] and V-Phaser 2 [35], were not employed, as these programs can only be used on the Linux platform.

QuRe [25] is a read graph-based, multi-threaded, and platform-independent software implemented in Java. This software requires a long-read (> 100 nt) dataset and a reference sequence for its input. Three calculation steps constitute this software; *k*-mer-based mapping, reconstruction of viral quasispecies sequences and their relative abundances, and a built-in Poisson error correction algorithm, which may also reduce NGS-derived artifacts. In this study, homopolymeric and non-homopolymeric error rate parameters were set to be 0.001, a value taken from a previously reported MiSeq error rate ($\sim$0.001) [19]. A post-reconstruction probabilistic clustering step was omitted. All calculations were iterated 1,000 times. For reference, either the core region or the NS3 protease region of either H77 (GenBank AF01175) or JFH1 (GenBank AB047639) sequence was used. The variant composition in each dataset was reconstructed and output as paired information of sequences and relative abundances.

QuasiRecomb is another QSR software implemented in Java, employing a strategy of probabilistic inference [26]. QuasiRecomb implements a hidden-Markov model for maximum a posteriori (MAP) parameter estimation, automatic model selection and prediction of the quasispecies distribution. It does not require prespecified references, but instead, mapped read set as an input. In this study, BAM files of mapping results generated from Geneious were used. QuasiRecomb also implements many option commands allowing flexible analysis. This time, the flag '-conservative' was not employed because our interest was on minor variants. Either the core or the NS3 protease region was specified using the '-r' command. The variant compositions were reconstructed as with QuRe.

Genotyping of reconstructed variants

Reconstructed variant sequences were aligned with the Los Alamos genotype reference sequence of either the core or the NS3 protease region using MAFFT [32], and phylogenetic trees were constructed using FastTree [36], both of which tools are implemented as Geneious plugins. Patristic distance matrix was calculated using Geneious from the resultant phylogenetic trees. Each reconstructed sequence was compared with all of the reference sequences, and intra-subtype average patristic distances were calculated using an in-house script. The geno/subtype minimizing the average distance was considered the geno/subtype of the reconstructed sequence.

RAV screening in NS3 protease region

Simeprevir is a noncovalent, macrocyclic NS3 protease inhibitor [37] and has been proven to be effective in combination with peg-IFN plus ribavirin [38–41] and an IFN-free regimen with sofosbuvir [10]. Despite its efficacy and the mildness of its side effects, there are several RAVs; the amino acid substitutions at V36, F43, Q80, S122, S138, R155, A156, V158, D168 and V170, have been reported to confer resistance against simeprevir [42,43]. Considering its clinical significance, RAVs associated with resistance against simeprevir and relevant DAAs were chosen for screening in this study.

Reconstructed variant sequences were aligned with the NS3 reference sequence using MAFFT, and further codon-aligned and translated using the Codon Alignment v1.1.0 web tool

(http://hcv.lanl.gov/content/sequence/CodonAlign/codonalign.html). After gaps were removed manually, relevant amino acid positions were scrutinized using in-house scripts, and the relative abundance of each RAV was calculated.

To assess the performance of QSR-based RAV screening, the SNV-based inference of RAVs was also attempted. BAM-formatted mapping files were used as inputs for the R package deepSNV [44], and SNV frequencies were estimated with the parameters 'sig.level' = 0.001 and 'adjust.method' = "BH". As a control counterpart for deepSNV calculation, the MiSeq sequencing data from *in vitro* transcribed control HCV RNA was used.

Simulation experiments of quasispecies reconstruction

To evaluate the performance of QuRe and QuasiRecomb, *in silico* simulation experiments were carried out. First, MiSeq sequencing files were obtained from three clinical specimens, in which different dominant Gts and amino acid substitutions at NS3 Q80 and/or S122 (Gt1b and Q80K + S122S, Gt1b and Q80Q + S122G, and, Gt2a and Q80G+S122K) were preliminarily identified. Next, mapping was performed, and reads that did not match the dominant substitution were removed. Finally, reads were randomly retrieved from each dataset according to prespecified ratio (see S2 Table) and combined *in silico* into one sequence set. Resultant datasets represent hypothetical quasispecies mixtures of different prespecified relative abundances. In this way, simulation experiments could be performed with sequencing error rates, read length distributions and other characteristics almost the same as the actual NGS. QSR, genotyping and RAV screening were performed as described above. True positives (TPs) indicate the existence of Gts or RAVs specified for simulation, and false negatives (FNs) indicate the failure to detect them. False positives (FPs) indicate the incorrect detection of unintended Gts or RAVs. Sensitivity (Sn) was calculated as the ratio of the number of TPs to the sum of the numbers of TPs and FNs; positive predictive value (PPV) was defined as the ratio of the number of TPs to the sum of the numbers of TPs and FPs.

Integrated analysis of the association of genotype and RAV for reconstructed quasispecies sequences

Genotyping and RAV screening were carried out for all reconstructed quasispecies sequences as discussed above. Results were then clustered according to (1) the QSR program used, (2) the sample ID, and (3) genotype. If any cluster contained at least one sequence having a specific RAV, the cluster was considered positive for that RAV. In this way, the following attributes were allocated to every cluster: name of QSR software, sample ID, status of HIV coinfection, history of blood exposure (BLx), genotype, and presence or absence of each RAV.

Using this data matrix, univariate and multivariate analyses were conducted to find nominal factors associated with specific RAVs. For univariate analysis, Fisher's exact test was conducted for each RAV. Significance level was not corrected for multiple testing, and a cut-off threshold was set at an unadjusted p-value of < 0.05 for the screening purpose. For multivariate analyses, logistic regression analyses were performed. Significantly associated Gt factors for each RAV were determined by backward stepwise selection with the cut-off threshold of adjusted p-value being less than 0.05. In the logistic regression analysis, p-values were corrected by Bonferroni's method, i.e., multiplied by the number of RAVs analyzed.

Results

Characterization of Illumina MiSeq NGS reads

The goal of our study was to simultaneously determine the composition of dominant and minor Gts, abundant and low-frequency RAVs, and characteristic combinations of Gts and RAVs from a clinical specimen from an HCV-infected patient. Therefore, we developed an in-house pipeline consisting of (1) NGS data generation, (2) NGS data cleaning, (3) QSR, (4) genotyping and (5) RAV screening of reconstructed sets of sequences, and (6) integration of Gts and RAVs determined from previous analyses of each reconstructed quasispecies.

For amplification, RT-PCR was performed using an in-house set of primers (see S1 Table), and the amplicon was excised from agarose gel, purified and analyzed using Illumina MiSeq. From 21 clinical samples, 14,558,762 sequences were obtained after removing low-quality reads and contaminating reads. The length distribution of adaptor-trimmed insert sequences is shown in Fig. 2A; the average length was 194.2 and the standard deviation was 61.0, with the minimum read length of 50 and the maximum of 301. Quality trimming was performed with a threshold of quality score < 20 for each read. The proportion of reads with the least quality score > 30 was 98.0%. Mapping to the HCV H77 sequence was carried out using the Geneious software to confirm uniform coverage (30864 ± 9619 as mean ± s.d.) throughout the amplified region (Fig. 2B).

Next, we estimated the rate of artificial nucleotide substitutions using control RNA (see Methods). The result of SNV screening by deepSNV demonstrated 93 out of 452 nucleotide positions in the HCV core region (463–914 in the genome of the H77 isolate) at the relative abundance range of 0.0145 ± 0.0691 (mean ± s.d.), and only two out of 600 in the upstream region of NS3 protease (3420–4019 in the H77 genome) with their relative abundances of 0.0174 and 0.0298. QSR-based genotyping resulted in Gt1b at an abundance of 1.00. RAV screening revealed no artificial RAVs. S122A was found in one of the duplicates at an abundance of 0.00032, although this variant does not confer resistance.

Characterization of QSR-based genotyping with simulated datasets

To examine the feasibility of performing QSR on our NGS datasets, we first checked the distributions of nucleotide mismatches between HCV reference sequences. S1 Fig. shows the distributions of SNV-to-SNV nucleotide distances in the core region (base position 463–914 in the genome of the H77 isolate) and the NS3 protease region (3420–4019 in the H77 genome) using HCV reference sequences retrieved from the Los Alamos HCV sequence database. The SNV-to-SNV intervals were significantly shorter than the NGS read length (Mann-Whitney one-tailed tests, $p < 10^{-10}$ in all subgroups shown in S1 Fig.).

Assuming the feasibility of performing QSR on the obtained NGS datasets, we then planned *in silico* simulation experiments with real NGS datasets obtained from clinical specimens. Three samples from HCV/HIV coinfected patients possessing dominant Gt and amino acid substitutions at NS3 Q80 and/or S122 (Gt1b and Q80K + S122S, Gt1b and Q80Q + S122G and Gt2a and Q80G+S122K, respectively) were selected (namely, 'HCVHIV04', 'HCVHIV05' and 'HCVHIV06'). NGS read datasets were fabricated by randomly taking reads from three selected sources (see S2 Table for detailed simulation parameters), and QSRs were performed. The genotyping results are summarized in Fig. 3. As for the genotyping of the core region, when only Gts observed commonly in the QSR results of QuRe and QuasiRecomb were retained, expected Gts (Gt1b and Gt2a) were detected under all simulation conditions, whereas no unexpected Gts were retained (Fig. 3A). In NS3, Gt2a (minor Gt) was overlooked in four simulations, in all cases of which the parameter of the total read count was set as low (L). When each Gt observed

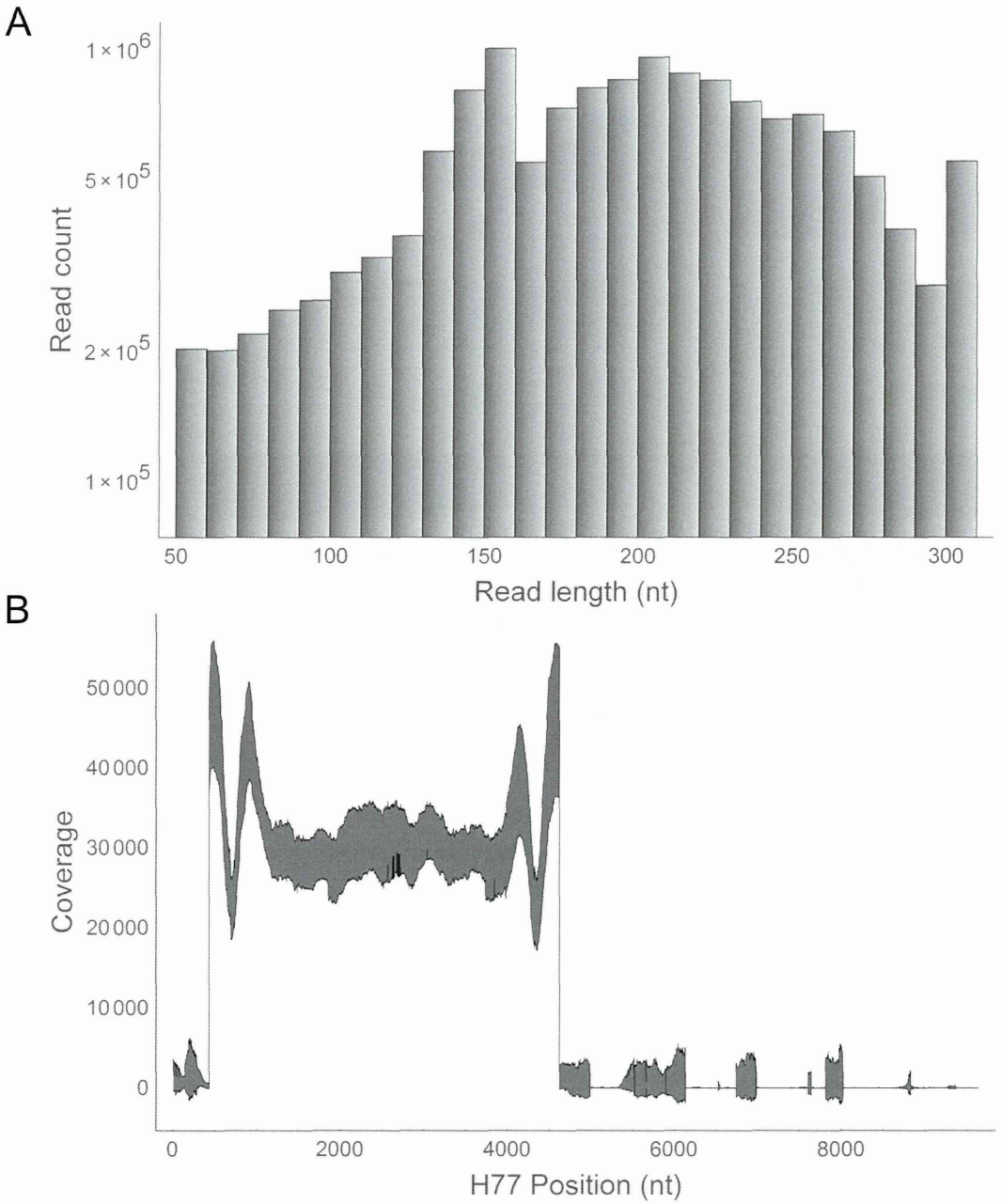

Fig 2. Characterization of Illumina MiSeq sequencing. (A) Read length histogram of all insert sequences of all clinical samples (n = 21). Insert sequences were adaptor-trimmed in advance. (B) Coverage plot showing the 95% confidence intervals of the coverages at all nucleotide positions calculated from all sequence datasets of clinical samples (n = 21). The core region spans from base positions 342 to 914, and the NS3 region spans from 3420 to 5312.

doi:10.1371/journal.pone.0119145.g002

at least once in the results of either QuRe or QuasiRecomb was retained, unexpected Gts (Gt1a, 2b, and 2k) appeared in the genotyping results of both the core and NS3 (Fig. 3B and 2D, respectively). Erroneously assigned Gts, mainly derived from QuasiRecomb data (data not shown), were equally distant from either Gt1b or Gt2a (S2 Fig.). Seven false-positive Gts were

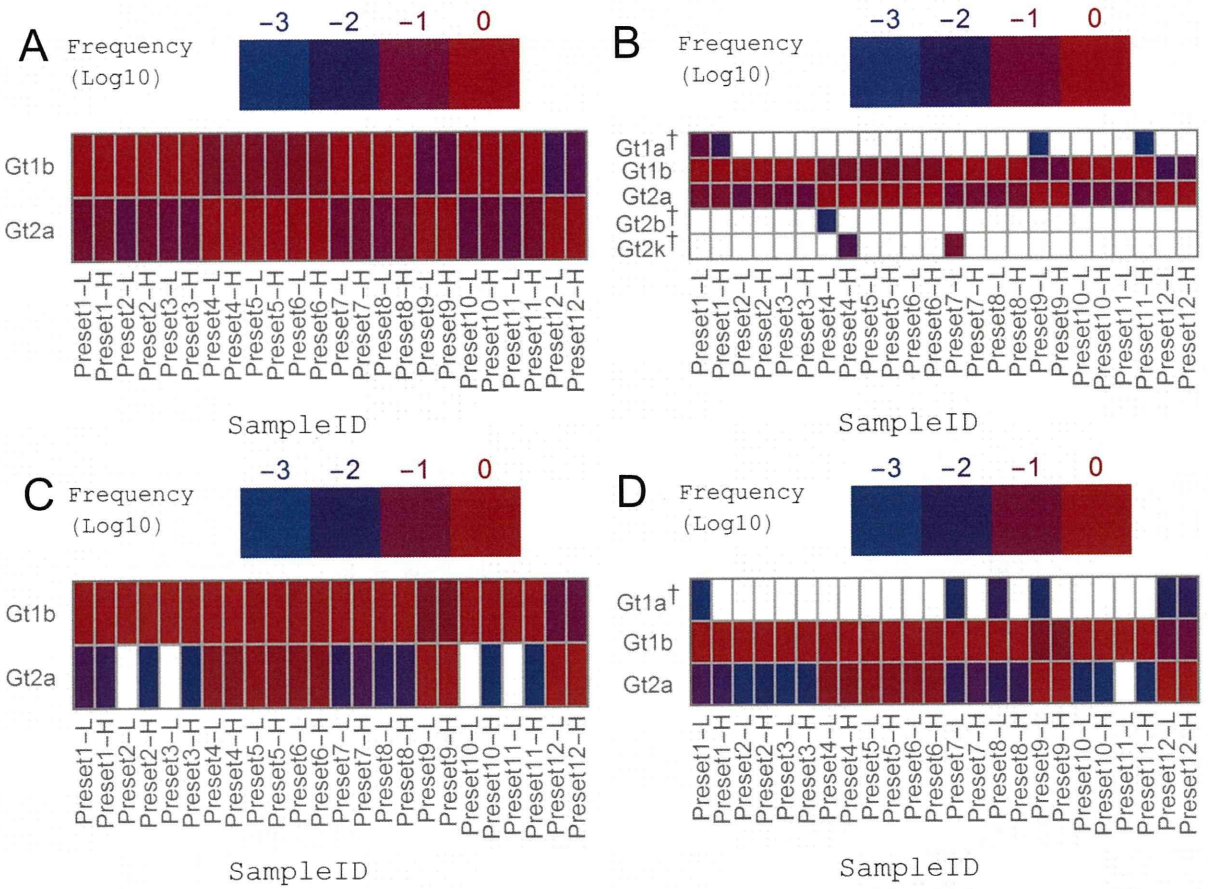

Fig 3. Combination of different QSRs can reduce false-negative genotypes and false-positive genotypes. Simulated datasets were used for QSR calculation followed by genotype (Gt) assignment using either QuRe (JFH1 was used as a reference) or QuasiRecomb. The x-axis labels denote the simulation settings of the preset ratio of relative abundance of intended genotypes (e.g., Gt1b: Gt1b: Gt2a = 95: 95: 5 in the Preset 1 dataset) and the total number of reads (L denoting 30,000 reads, and H denoting 100,000 reads). See S2 Table for all simulation conditions. The y-axis labels are the observed Gts. False-positive Gts (Gts other than Gt1b and Gt2a) are labeled with a dagger (†). (A) Gts observed in both QuRe and QuasiRecomb reconstructions targeting the core region. (B) Gts observed at least once in either QuRe or QuasiRecomb reconstruction targeting the core region. (C) Gts observed in both QuRe and QuasiRecomb reconstructions targeting the NS3 protease region. (D) Gts observed at least once in either QuRe or QuasiRecomb reconstruction targeting the NS3 protease region. From the comparison of the results of QuRe and QuasiRecomb, higher abundances were always selected. The threshold was set at a frequency of 0.001.

doi:10.1371/journal.pone.0119145.g003

detected in the QSR of the core (Fig. 3B), whereas six false-positive Gts and one false-negative Gt were detected in the QSR of the NS3 protease region (Fig. 3D). Sns and PPVs are summarized in Table 2.

To characterize the quantitative reliability of this genotyping approach, the estimated relative abundance of each Gt was compared with the preset abundance for simulation (S3 Fig.). As for the core region, QuRe reconstructed both dominant and minor Gts quantitatively in all of the simulation conditions tested. Although QuasiRecomb also successfully reconstructed both dominant and minor variants, the abundances of minor variants were more likely estimated to be larger than the preset values, 0.010 and 0.050. QuasiRecomb reconstructed three false-positive Gts at the frequency range of 0.0035 to 0.0497 (S3A Fig.). QuRe also generated three false- positive Gts but their estimated abundances were at the maximum of 0.0013, much smaller than the values of those incorrectly reconstructed by QuasiRecomb (S3B Fig.). In the

Table 2. Properties of QSR-based genotyping.

	Genotyping Method	TP [c]	FP [d]	FN [e]	Sn [f]	PPV [g]
Core	QuRe AND QuasiRecomb [a]	48	0	0	100.0%	100.0%
	QuRe OR QuasiRecomb [b]	48	7	0	100.0%	87.3%
NS3	QuRe AND QuasiRecomb [a]	44	0	4	100.0%	100.0%
	QuRe OR QuasiRecomb [b]	47	6	1	100.0%	88.7%

[a] QuRe AND QuasiRecomb: Reproducibly detected by both QuRe and QuasiRecomb

[b] QuRe OR QuasiRecomb: Detected at least once by either QuRe or QuasiRecomb

[c] TP: The number of true positives (expected and correctly detected cases)

[d] FP: The number of false positives (unintended but incorrectly detected cases)

[e] FN: The number of false negatives (expected but incorrectly overlooked cases)

[f] Sn: Sensitivity = TP / (TP + FN)

[g] PPV: Positive predictive value = TP / (TP + FP)

doi:10.1371/journal.pone.0119145.t002

NS3 protease region, QuRe generated no false-positive Gts but one false-negative Gt under the simulating conditions of a low (L) read count and a preset abundance of 0.010 (S3C Fig.), whereas QuasiRecomb yielded not only one false-negative Gt but also six false-positive Gts at the estimated abundances ranging from 0.0025 to 0.0155 (S3D Fig.).

Characterization of QSR-based RAV screening using simulated datasets

The in-house bioinformatics pipelines for the detection of RAVs in the NS3 protease region were tested using the same simulated datasets discussed above (Fig. 4). When only RAVs reproducibly detected from the results of QuRe and QuasiRecomb were retained, expected RAVs (Q80K, S122G, and Q80G+S122K) were detected with an overall Sn and PPV of 80.6% (58/72) and 100.0% (58/58), respectively (Fig. 4A and Table 3). All of the unexpected RAVs uniquely detected by either QuRe or QuasiRecomb were automatically removed through the consensus-making step. In contrast, when all RAVs observed at least once were kept, Sn increased to 98.6% (71/72) whereas PPV dropped to 41.8% (71/170), apparently owing to the large number of false-positive RAVs (Fig. 4B). We then traced the origin of those false-positive Gts; QuRe had a Sn of 86.1% (62/72) and a PPV of 88.6% (62/70), whereas QuasiRecomb had a slightly higher Sn (93.1%, 67/72) but a much lower PPV (42.9%, 67/156). All the Sns and PPVs are summarized in Table 3.

QSR-based genotyping with clinical samples

We then attempted to apply our pipelines to the analyses of clinical samples. The genotyping results of 21 HCV-infected patients are summarized in Fig. 5. Because the genotyping strategy of using the core region and taking the consensus of QuRe and QuasiRecomb outperformed other options in the simulation experiments discussed above, we first focused on this strategy (Fig. 5A). Notably, in eight out of 11 HCV/HIV coinfected patients, the dominant Gts were non-Gt1b (6 Gt1a, one Gt2a and one Gt2b), whereas in all but 'HCVmono28' HCV monoinfected patients, Gt1b was dominant. Gt1a infection was dominant only in HCV/HIV coinfected hemophiliacs (6/11 vs 0/10, $p = 0.0124$). Further genotype analysis indicated the presence of multi-geno/subtype overlapping infection in 7 out of 11 HCV/HIV coinfected hemophiliacs and 4 out of 5 HCV monoinfected patients with a history of whole-blood

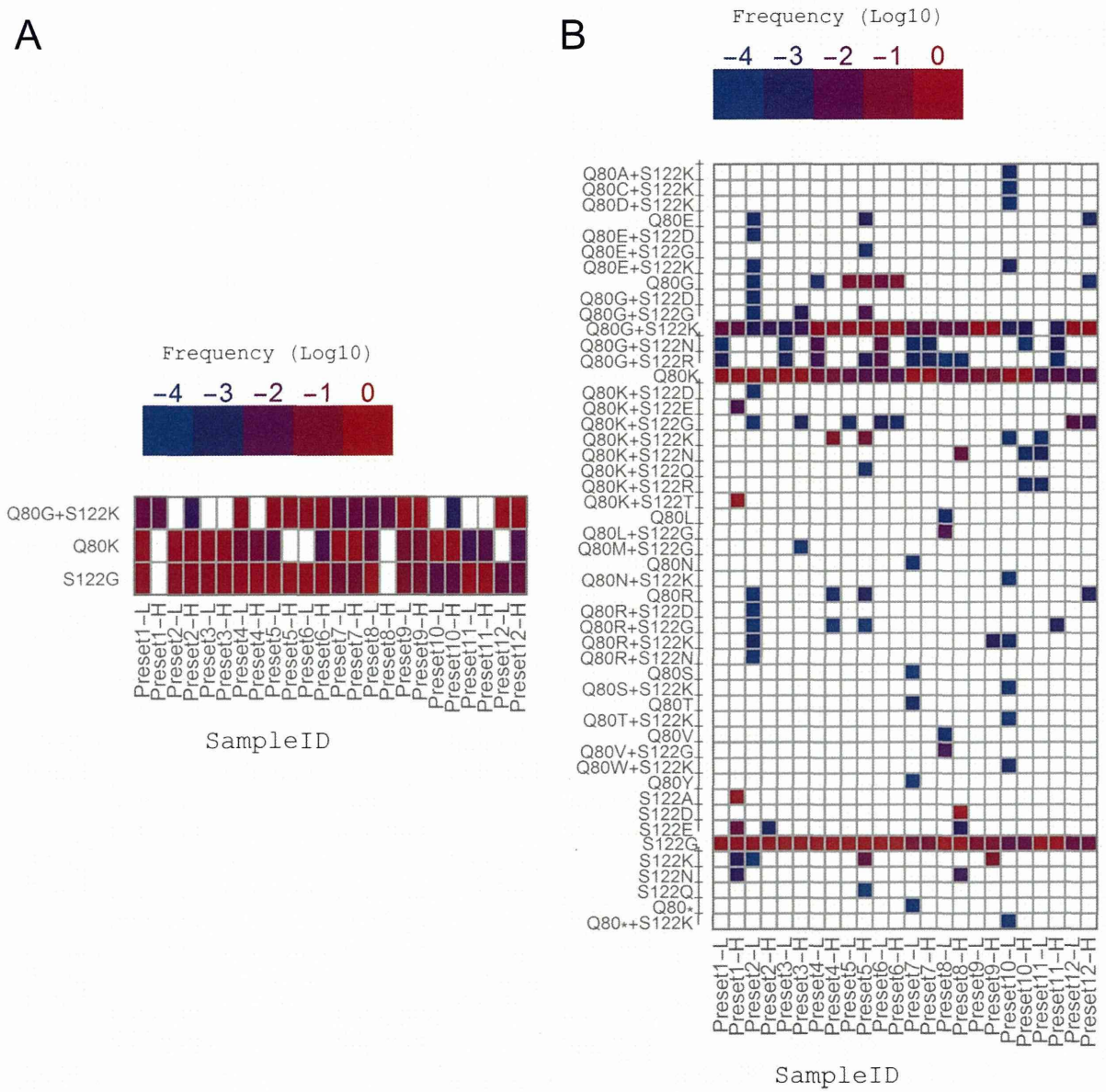

Fig 4. Low-frequency false-positive RAVs can be effectively removed by filtering out variants not reproducibly detected in different QSRs. Simulated datasets were used for QSR calculation followed by the screening of RAVs using either QuRe (JFH1 was used as a reference) or QuasiRecomb. The x-axis labels denote the simulation settings of preset ratio of relative abundance of intended RAVs (e.g., Q80G+S122K: Q80K: S122G = 5: 60: 35 in the Preset 1 dataset) and total number of reads (L denoting 30,000 reads, and H denoting 100,000 reads). See S2 Table for all simulation conditions. The y-axis labels are the observed RAVs. False-positive RAVs (RAVs other than Q80G+S122K, Q80K and S122G) are labeled a dagger (†). (A) RAVs observed in both QuRe and QuasiRecomb reconstructions. (B) RAVs observed at least once in either QuRe or QuasiRecomb reconstruction. From this comparison between the results of QuRe and QuasiRecomb, larger abundances were always selected. The threshold was set at a frequency of 0.0001.

doi:10.1371/journal.pone.0119145.g004

transfusion, whereas none among 5 HCV monoinfected patients without a history of whole-blood transfusion (Fig. 5A). When employing the strategy of incorporating every Gt observed, multi-geno/subtype infection was suspected in 10 out of 11 HCV/HIV coinfected hemophiliacs and 4/5 HCV monoinfected cases with a history of blood transfusion, in an apparent contrast