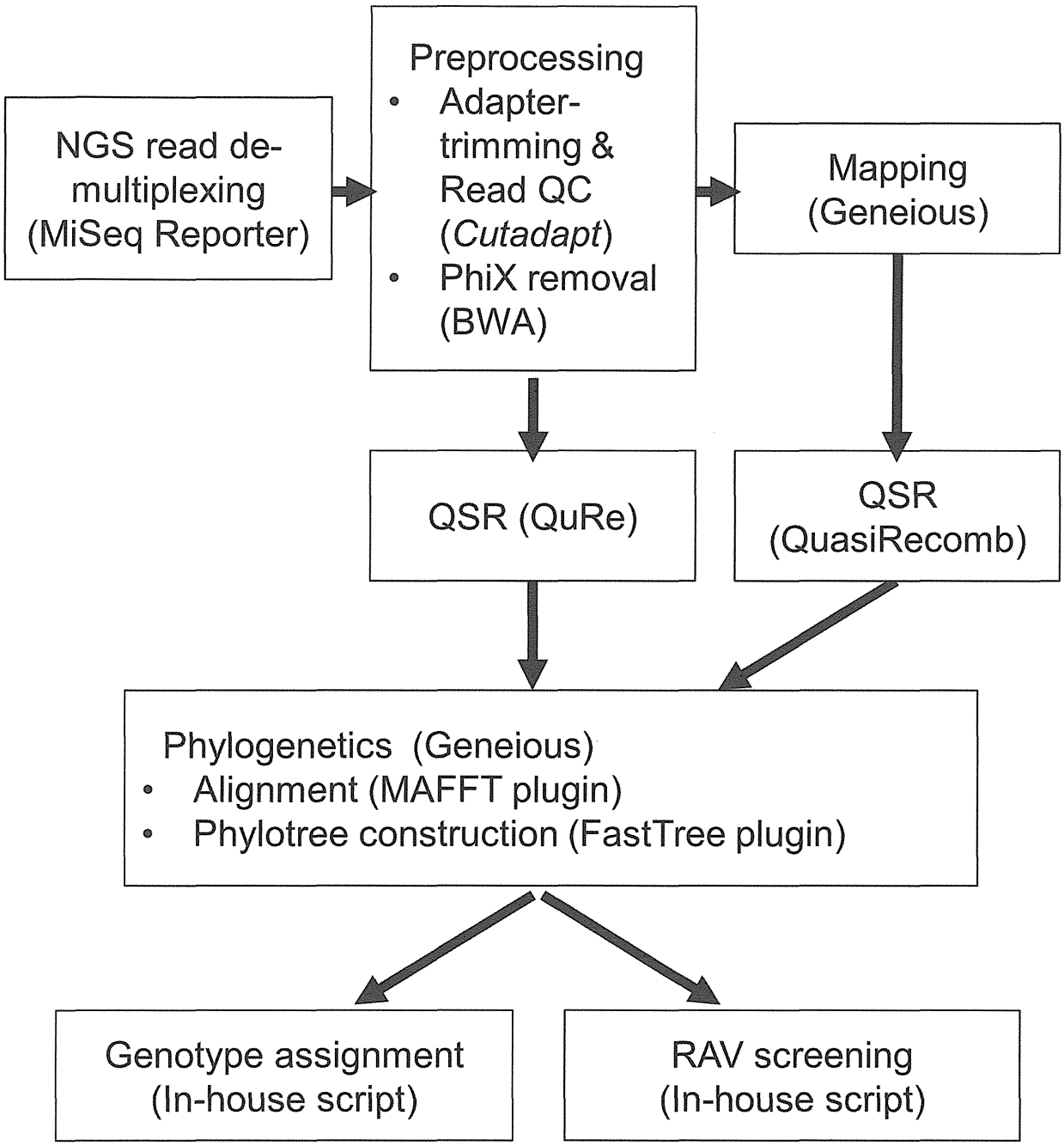

Fig 1. A flowchart of bioinformatics employed in this study.

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

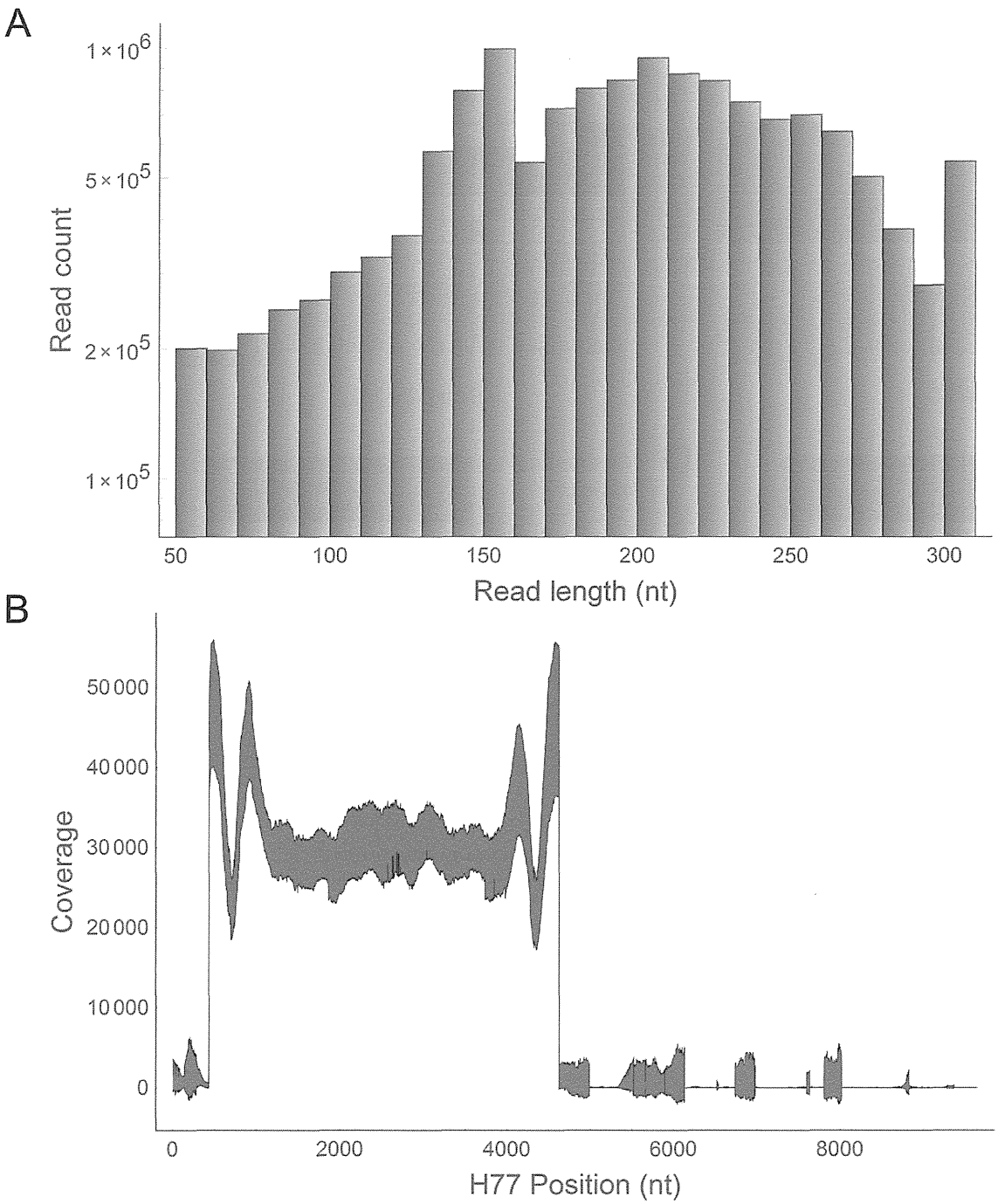

Fig 2. Characterization of Illumina MiSeq sequencing. (A) Read length histogram of all insert sequences of all clinical samples (n = 21). Insert sequences were adaptor-trimmed in advance. (B) Coverage plot showing the 95% confidence intervals of the coverages at all nucleotide positions calculated from 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

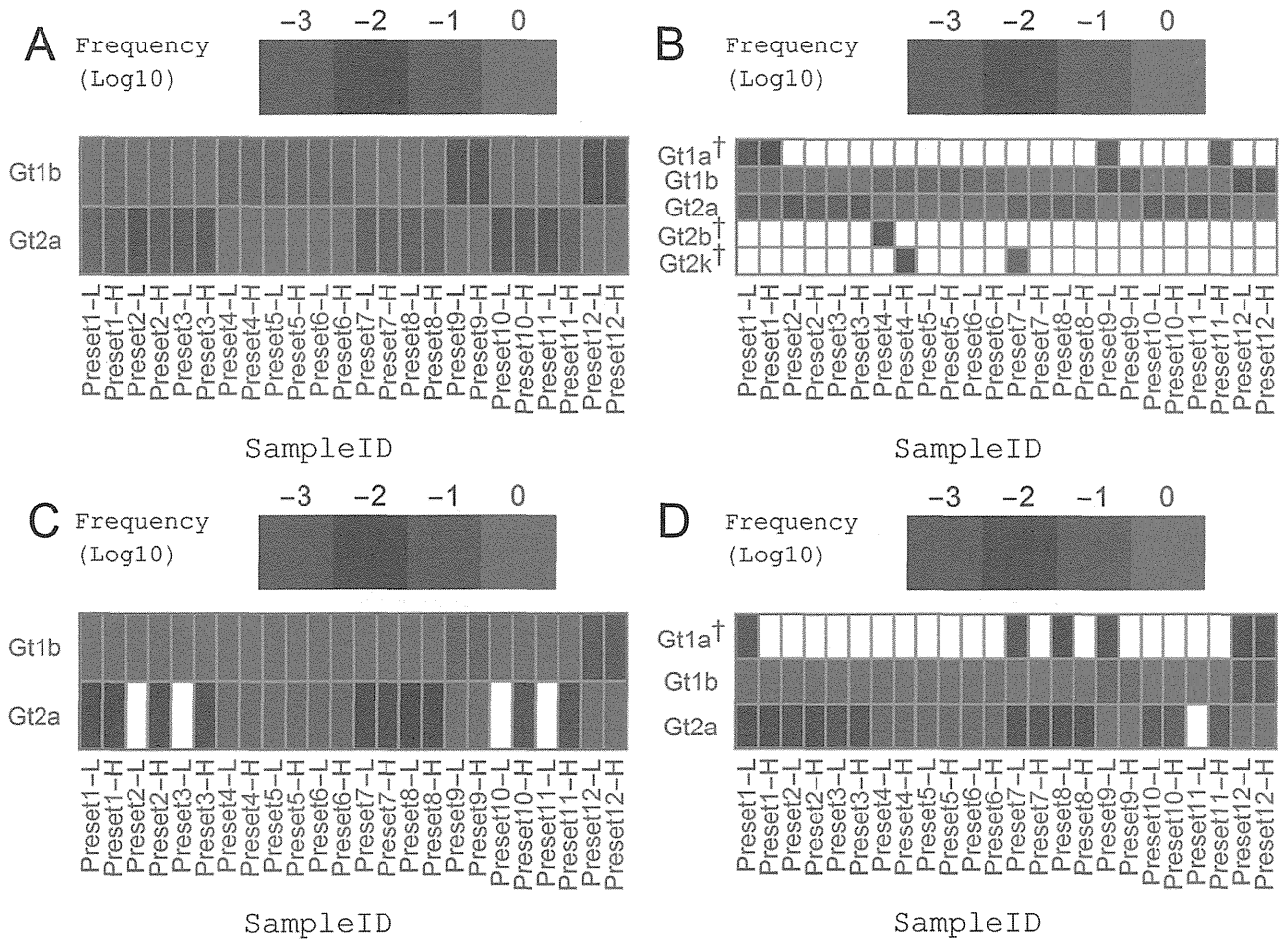

Fig 3. Combination of different QSRs can reduce false-negative genotypes and false-positive genotypes. Simulated datasets were used for QSR calculation followed by genotype (Gt) assignment using either QuRe (JFH1 was used as a reference) or QuasiRecomb. The x-axis labels denote the simulation settings 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

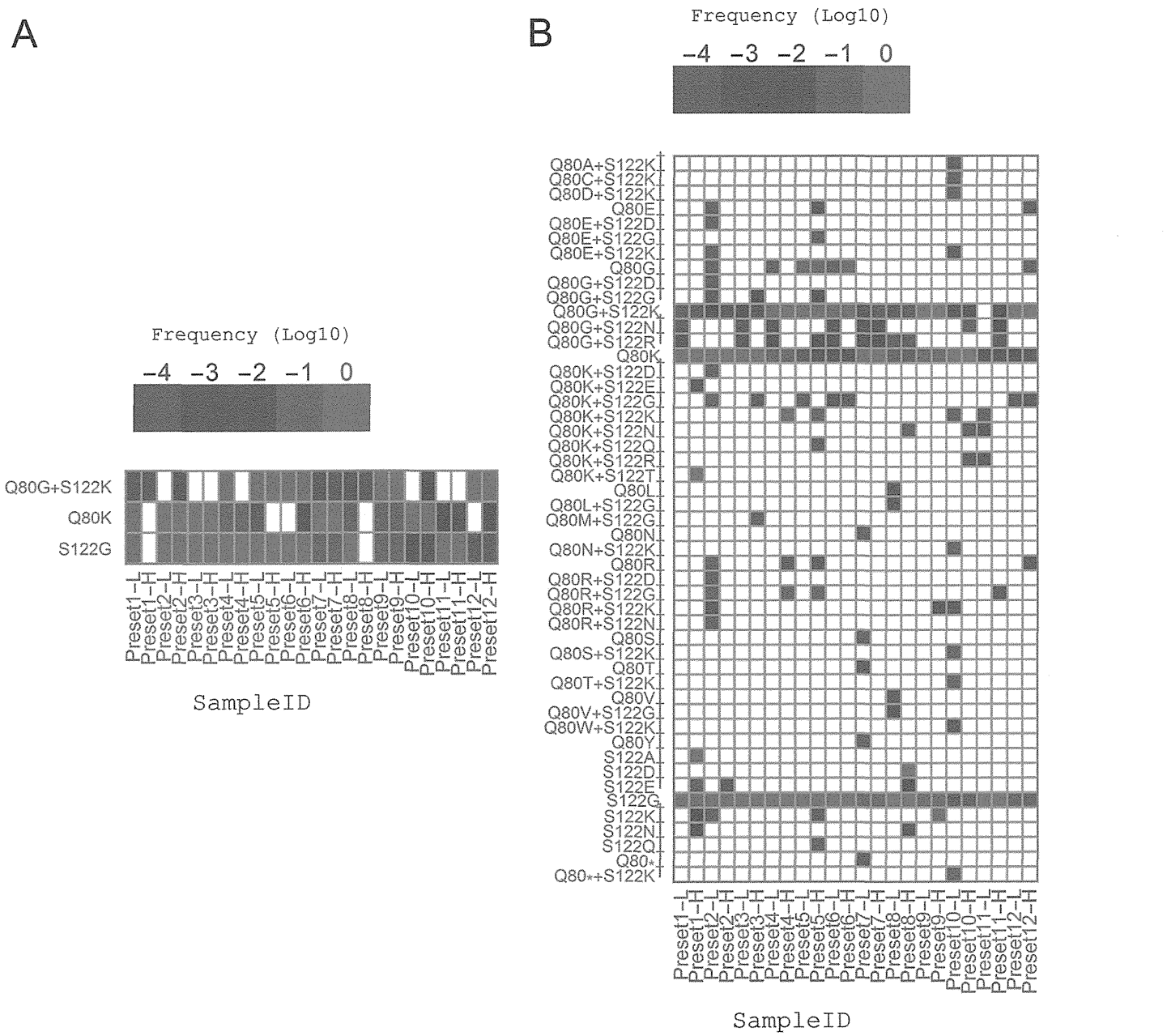

Fig 4. Low-frequency false-positive RAVs can be effectively removed by filtering out variants not reproducibly detected in different QSRs. Simulated datasets were used for QSR calculation followed by the screening of RAVs using either QuRe (JFH1 was used as a reference) or QuasiRecomb. The x-axis labels 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

Table 3. Properties of QSR-based RAV screening.

	RAV Screening Method	TP [c]	FP [d]	FN [e]	Sn [f]	PPV [g]
NS3	QuRe AND QuasiRecomb [a]	58	0	14	80.6%	100.0%
	QuRe	62	8	10	86.1%	88.6%
	QuasiRecomb	67	89	5	93.1%	42.9%
	QuRe OR QuasiRecomb [b]	71	99	1	98.6%	41.8%

[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.t003

with those cases without a history of blood transfusion (Fig. 5B). The prevalence of multi-geno/subtype overlapping infection was significantly higher in a population with any history of exposure to blood (BLx) ($p = 0.0124$ and $p = 0.0010$ for the genotyping pipeline with or without consensus-based selection, respectively; Fig. 5A and 4B). When the NS3 protease region was used for genotyping, the most dominant genotype estimated in each subject was in good agreement with the genotyping results of the core region. However, overlapping infection was detected in only 3/11 HCV/HIV coinfected patients and 1/10 HCV monoinfected patients when consensus was taken between the results of QuRe and QuasiRecomb (Fig. 5C), and 9/11 HCV/HIV coinfected and 1/10 HCV monoinfected patients when all the genotypes detected by either QuRe or QuasiRecomb-based genotyping were included (Fig. 5D). It was notable that none of the 5 HCV monoinfected patients without BLx had overlapping infection detected by any of the genotyping strategies (Fig. 5).

QSR-based RAV screening using clinical samples

Considering the clinical significance of simeprevir, RAVs associated with resistance against simeprevir and relevant DAAs were chosen for subsequent analyses. Screening results are summarized in Fig. 6. Ten RAVs remained after removing disagreement between the QSR results of QuRe with the H77 sequence as the reference, QuRe with the JFH1 sequence as the reference, and QuasiRecomb. Eight of 10 variants were related to either Q80 or S122. No variants at positions R155, A156, V158, and D168 were definitively proven. It was notable that only 13 variants were detected using QuRe (S4 Fig.), whereas the total number markedly increased to 65 in the case of using QuasiRecomb (S5 Fig.). Q80K was detected in 4 out of 11 HCV/HIV co-infected hemophiliacs, whereas Q80R was detected in 1 out of 11 patients coinfected with HIV and HCV, and 7 out of 10 HCV monoinfected patients ($p = 0.0075$). A set of V36, Q80G, and either S122K or S122R was observed in patients 'HCVHIV06', 'HCVHIV15', and 'HCVmono28', all of whom had Gt2 as the dominant genotype. Low-frequency S122K and S122R were detected in one ('HCVmono15') and two ('HCVHIV16' and 'HCVmono28') cases, respectively.

Q80 and S122 have been associated with decreased viral sensitivity and treatment failure [42]. Therefore, we decided to focus on Q80K, Q80R, S122K, and S122R, all of which cause resistance against simeprevir with the fold change of more than 2 (considered moderate resistance) in Gt1a and Gt1b. After validating their existence by manually inspecting mapping data

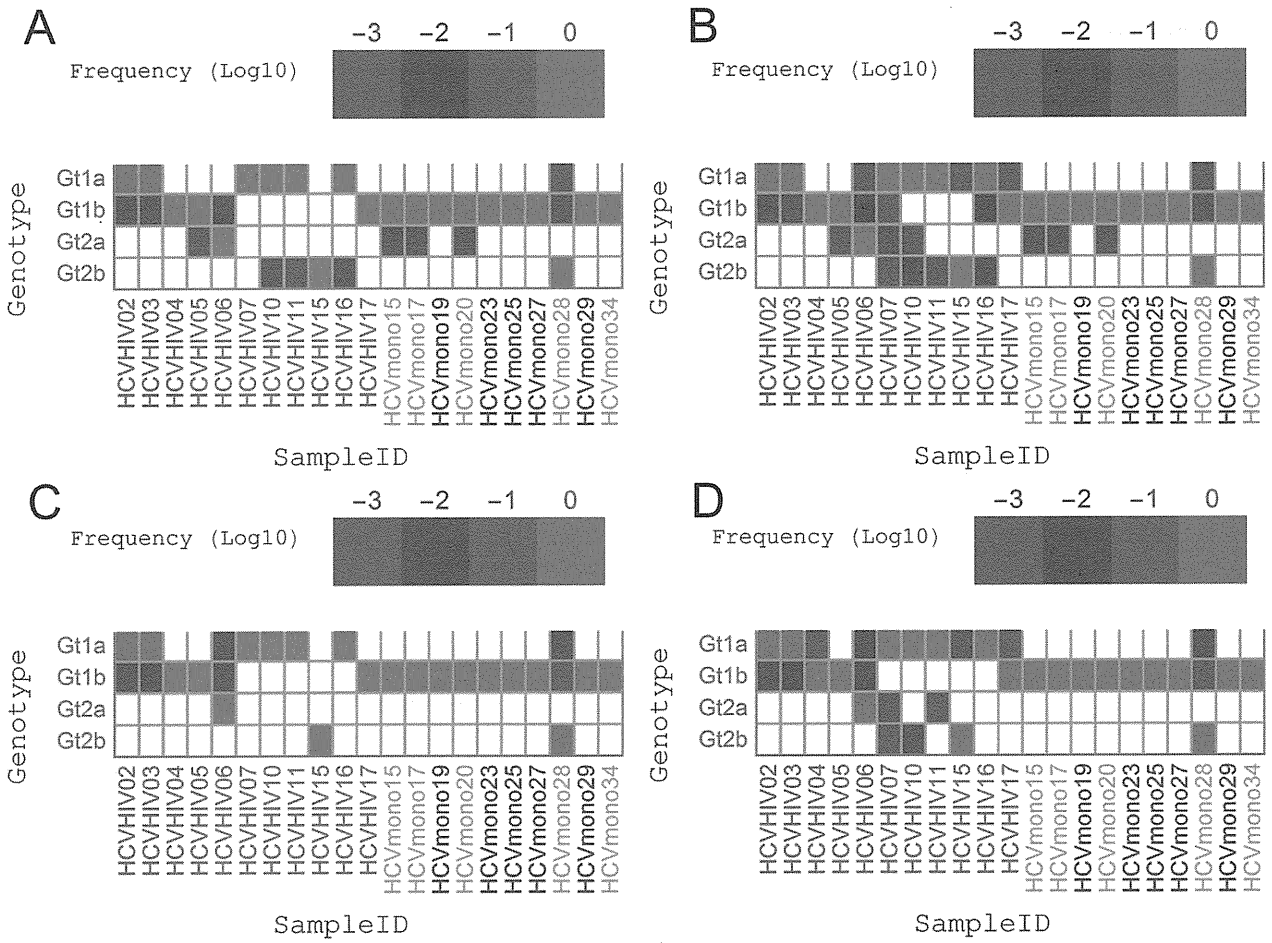

Fig 5. Prevalence of minor multigenotype infections in patients exposed to blood products. Relative abundances of minor genotypes (Gts) were estimated from the genotyping results of reconstructed quasispecies in each subject. The x-axis labels are sample IDs, colored on the basis of the patients' history of exposure to blood (see Table 1 for details). (A, C) Gts observed in both QuRe and QuasiRecomb reconstructions targeting (A) the core region and (C) the NS3 protease region. (B, D) Gts observed at least once in either QuRe or QuasiRecomb reconstruction targeting (B) the core region and (D) the NS3 protease region. From the comparison of the results of QuRe and QuasiRecomb, larger abundances were always selected. The threshold was set at a frequency of 0.001.

doi:10.1371/journal.pone.0119145.g005

(data not shown), we compared estimated abundances of each RAV by (1) QSR-based screening with consensus selection, and (2) SNV-based inference of RAVs (deepSNV), where the R package 'deepSNV' was used to estimate the frequencies of relevant SNVs. The results are shown in Fig. 7. Q80K was detected in 4 out of 11 HCV/HIV coinfected hemophiliacs but not in any of the 10 HCV monoinfected patients (Fig. 7A). Q80K was also detected by deepSNV in those 4 cases. However, there were 2 patients in whom Q80K was indirectly inferred on the basis of SNVs but not by the QSR-based screening (Fig. 7A). Q80R was detected in 1 out of 11 HCV/HIV co-infected and 7 out of 10 HCV monoinfected patients. The program deepSNV failed to detect Q80R in 5 out of 8 cases and incorrectly inferred its existence in one case, HCVmono28, wherein the reference codon was CAA, the corresponding variant codon was GGG, and the incorrectly inferred codon was CGA (the responsible SNV is underlined hereafter; Fig. 7B). As for S122 variants, S122K was detected in 'HCVHIV06' by QSR (Fig. 7C). The deepSNV-based screening failed to detect S122K (the corresponding codon was AAG) in 'HCVHIV06' and incorrectly interpreted the relevant SNVs as S122R (the reference codon was AGC and the incorrectly inferred codon was AGG) (Fig. 7C and 6D). In 'HCVHIV16', S122R

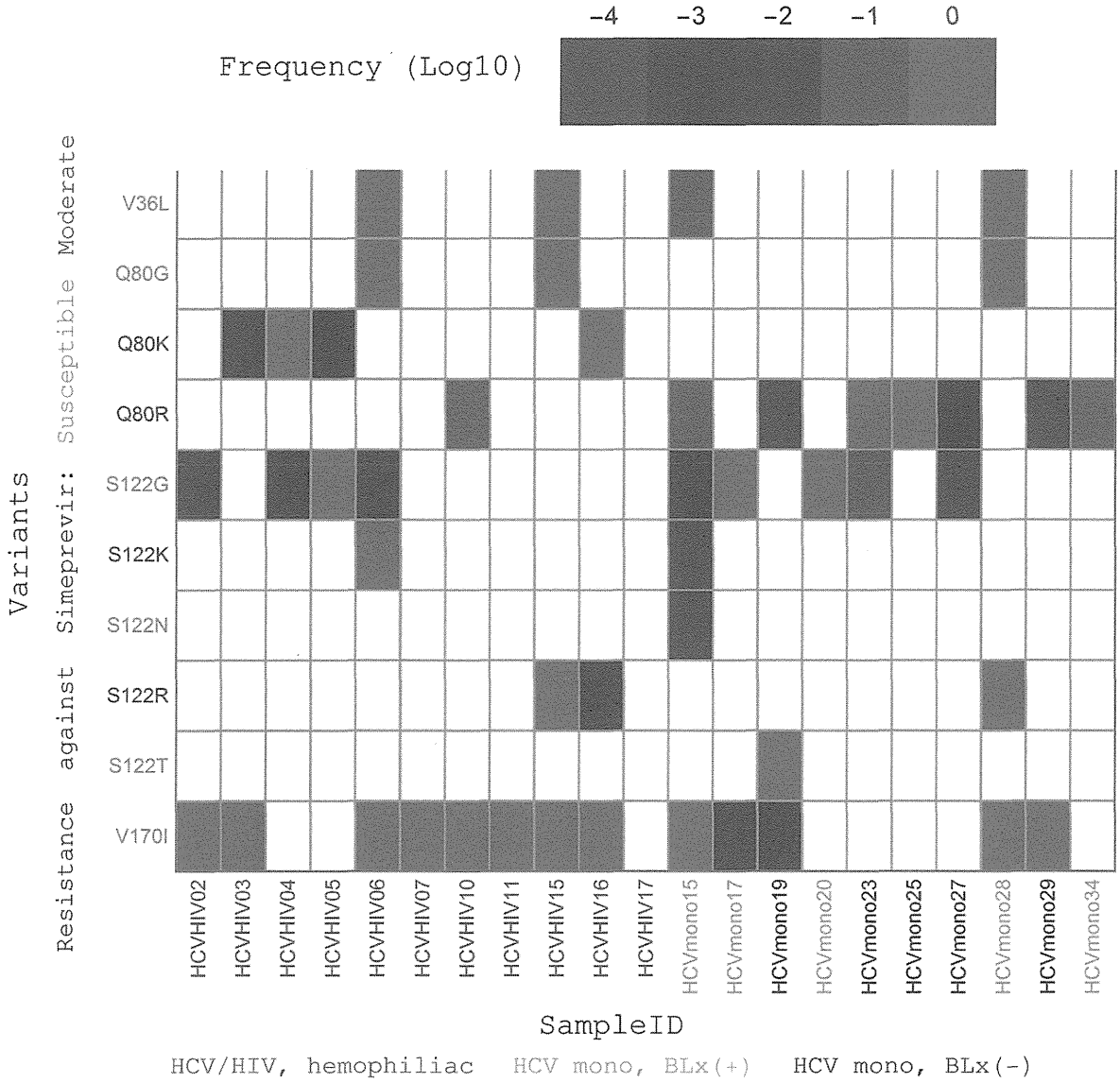

Fig 6. NS3 PI RAVs reproducibly detected by QSR-based screening. Relative abundances of resistance-associated variants (RAVs) in the NS3 protease region were estimated in each subject. Only RAVs reproducibly detected in the QSR results of QuRe and QuasiRecomb were retained. The x-axis labels are sample IDs, colored on the basis of their history of exposure to blood (see Table 1 for details). The y-axis RAV labels are colored on the basis of the effects of RAVs on simeprevir susceptibility: susceptible (FC < 2) substitutions are in cyan, moderately resistant substitutions (2 < FC < 50) in magenta. Highly resistant substitutions (FC > 50) were not detected. The threshold was set at a frequency of 0.0001.

doi:10.1371/journal.pone.0119145.g006

was detected by QSR but not by deepSNV-based screening. Mapping files were manually checked to confirm the correctness of RAVs detected by QSR (data not shown).

Integrated analysis of genotypes and RAVs

Because multi-geno/subtype overlapping infection was prevalent in this study cohort, there could be multiple cases wherein observed RAVs should be allocated to quasispecies of different genotypes. Therefore, we aimed at linking the Gts and RAVs. All reconstructed quasispecies sequences were clustered according to the sample ID and assigned genotype, and within each cluster, the presence or absence of RAVs was determined (see Methods). Univariate and

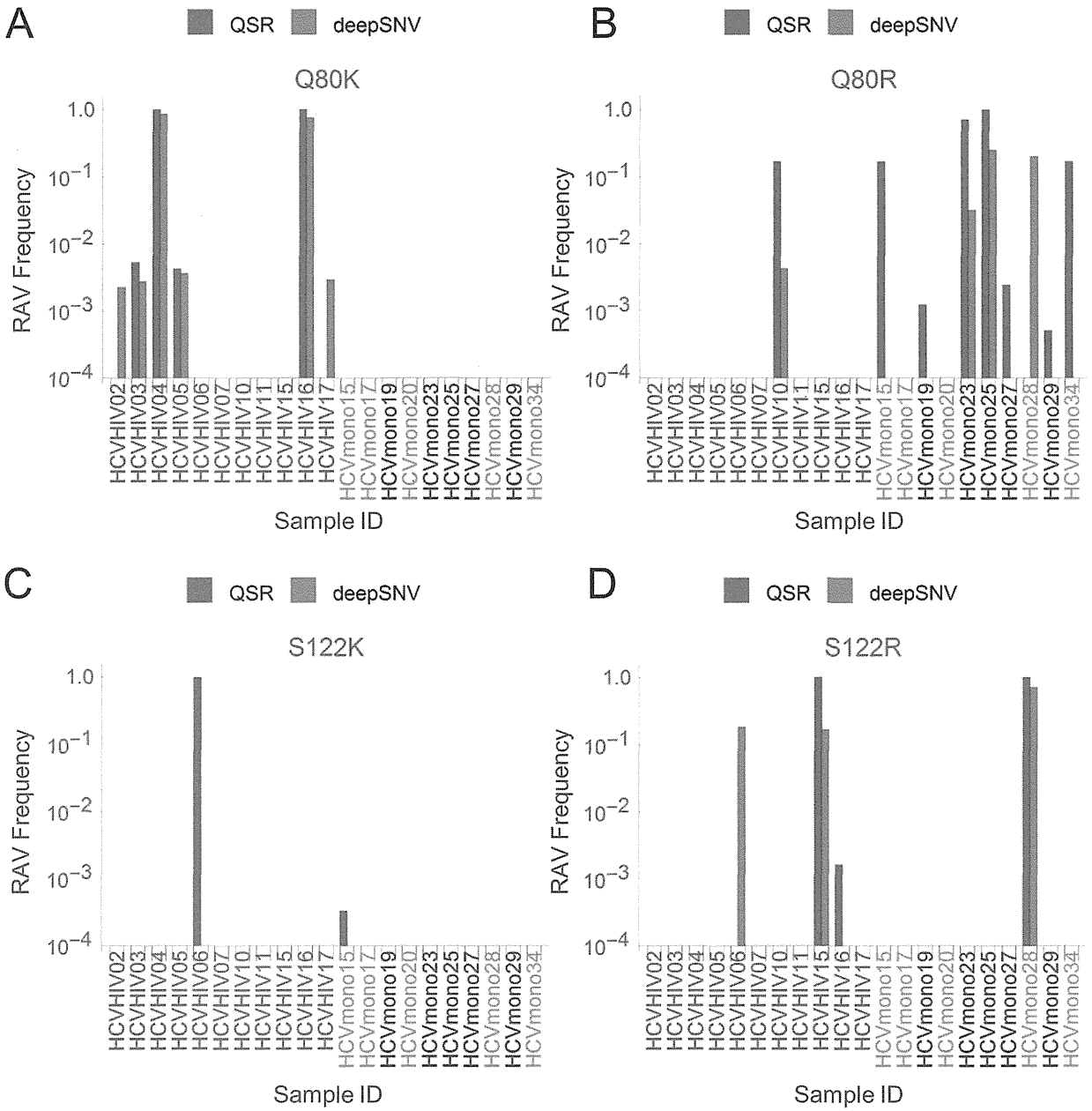

Fig 7. QSR-based RAV screening can detect amino acid changes derived from multiple nucleotide substitutions. The relative abundances of (A) Q80K, (B) Q80R, (C) S122K, and (D) S122R determined by QSR-based RAV screening and inference from SNVs detected by deepSNV (see Method).

doi:10.1371/journal.pone.0119145.g007

multivariate interdependence analyses for 94 clusters revealed that Q80K and Q80R were significantly associated with hemophilia and BLx (Tables 4 and 5). When focusing on HCV/HIV coinfected hemophiliacs, the association between Q80K and Gt1b was statistically significant [odds ratio (OR) = 13.4 (3.48–51.9), $p < 0.01$]. In contrast, when focusing on quasispecies reconstructed from HCV monoinfected patients, Q80K was inclined to have a certain association with Gt1a, although the p-value was not below the threshold of statistical significance after correction for multiple testing [OR = 70 (3.10–1580), uncorrected $p < 0.05$]. Univariate analysis indicated a negative correlation between hemophilia and the presence of Q80R [OR = 0.13,

Table 4. Univariate analysis of nominal factors associated with NS3 Q80K/R and S122K/R.

	Univariate					
RAV [a]	Odds ratio					
	Hemophilia [c]	BLx [d]	GT1a [b]	GT1b	GT2a	GT2b
Q80K	5.84 [††]	Inf [e, †]				
Q80R	0.13 [†††]	0	0.21 [††]	3.49 [††]		
S122K					42.0 [†††]	
S122R		Inf [†]		0.23 [††]		37.3 [†††]

[a] RAV: Resistance-associated variant

[b] Gt: genotype

[c] Hemophilia: Hemophilias with HCV/HIV coinfection and multiple exposures to unheated coagulation factor concentrates

[d] BLx: Exposure to unheated coagulation factor concentrates, or history of whole-blood transfusion

[e] Inf: Infinity

[†] Uncorrected $p < 0.05$

[††] uncorrected $p < 0.01$

[†††] uncorrected $p < 0.001$

doi:10.1371/journal.pone.0119145.t004

$p < 0.001$], although this was not proven by multivariate analysis. On the other hand, S122 variants were not statistically linked with either hemophilia or BLx. However, when the linkage of RAVs on the same reconstructed sequences was taken into consideration, the linked variants of S122K + Q80G were associated with Gt2a [OR = 174 (12.9–2350), $p < 0.05$], and those of S122R + V36L + Q80G were associated with Gt2b [OR = 145.0 (17.7–1190), $p < 0.001$].

We then attempted to further characterize the amino acid linkage, phylogenetics, and the distributions of estimated frequencies of quasispecies sequences having those variants. Q80K and Q80R were focused on in subsequent analyses as these variants were presumably linked with Gt1 infection, a principal target of many DAA therapies such as simeprevir. Quasispecies of Gt1a with Q80K, Gt1a with Q80R, Gt1b with Q80K, and Gt1b and Q80R were detected from combined QSR results of all samples (Fig. 8). A phylogenetic tree was constructed, which indicated a distinct subpopulation with each combination of genotype and Q80 substitution (Fig. 8A). Sequences assigned to each cluster were retrieved, aligned, and visualized as a sequence logo (Fig. 8B). Analyses of amino acids at positions 70–90 revealed that the sequences of Gt1b had Gt1b-specific amino acids, whereas the sequences of Gt1a had Gt1a-specific amino acids at positions 71,72, and 89 (V, I, and Q for Gt1a, and I, T, and P for Gt1b, respectively), regardless of the amino acid variant at position 80. However, Gt1a-Q80R sequences had V78, which is the characteristic of Gt1b. The codon usage patterns at position 80 also differed from one another (Fig. 8B and 8D). The most dominant codons were AAA(K), CGA(R), AAG(K) and CGG(R) and their relative frequencies were 99.4%, 84.0%, 97.9% and 98.1% for Gt1a-Q80K, Gt1a-Q80R, Gt1b-Q80K and Gt1b-Q80R, respectively. Fig. 8C shows the distributions of relative frequency per reconstructed quasispecies sequence having these Gt–RAV pairs. In every Gt–RAV pair, the relative frequencies ranged from ~0.01% to ~100% with no remarkable differences.

Furthermore, the Los Alamos HCV sequence database [31] was scrutinized to find previously registered cases of the combinations of Gt1a-Q80K, Gt1a-Q80R, Gt1b-Q80K, and Gt1b-Q80R. All the sequences containing the NS3 region were analyzed, and the registered sequences were binned by genotype and sampling country (S3 and S4 Tables, respectively). Surprisingly, there was only one sequence of Gt1b with Q80K, and there were only three cases of

Table 5. Multivariate analysis of nominal factors associated with NS3 Q80K/R and S122K/R.

	Multivariate					
RAV [a]	Odds ratio [95% confidence interval]					
	Hemophilia [c]	BLx [d]	GT1a [b]	GT1b	GT2a	GT2b
Q80K	13.6 [3.14–58.5] *		70 [3.10–1580] †,##	13.4 [3.48–51.9] **,#		
Q80R	0.13 [0.05–0.32] ***					
Q80R +V36L		0.02 [0.002–0.15] *				
S122K						
S122K +Q80G					174. [12.9–2350] *	
S122R			0.10 [0.02–0.54]†,##	0.02 [0.002–0.27] †,##		44.7 [5.11–390] *
S122R +V36L						624 [35.5–11000] *
S122R +Q80G						273 [21.9–3400] *
S122R +V36L +Q80G						145 [17.7–1190] ***

[a] RAV: Resistance-associated variant

[b] Gt: genotype

[c] Hemophilia: Hemophiliacs with HCV/HIV coinfection and multiple exposures to unheated coagulation factor concentrates

[d] BLx: Exposure to unheated coagulation factor concentrates, or experience of whole-blood transfusion

† Uncorrected $p < .05$

* $p < .05$

** $p < .01$

*** $p < .001$

Logistic analysis with only 'Hemophilia' (+) clusters included

Logistic analysis with only 'Hemophilia' (–) clusters included

doi:10.1371/journal.pone.0119145.t005

Gt1a with Q80R (S2 Table). Most reports of sequences with Q80K or Q80R were from the United States, and none were previously registered in Japan (S3 Table).

Discussion

In this study, we developed and characterized the Illumina MiSeq NGS sequencing system coupled with a novel, accurate, and high-throughput bioinformatics pipeline involving quasispecies reconstruction (QSR), genotype (Gt) assignment, screening of resistance-associated amino acid variants (RAVs), and integrative analysis of the association between Gts and RAVs.

Our approach has several novelties compared with those used in previous studies. First, many previous studies have used Roche pyrosequencing-based NGS, not Illumina's flow-cell-based sequencer. Loman et al. characterized the performance of several bench-top NGS sequences, concluding that Illumina MiSeq has the highest read generation capability with the lowest frequency of sequence errors, particularly indels [19]. Because indels and other sequencing errors could result in false-positive low-frequency RAVs, a high read coverage and a low error rate would be preferred in viral research. Next, we employed the QSR technique to (1) eliminate sequencing errors through the reconstruction step, and (2) obtain sets of haplotypes spanning the genome region of interest (mimicking cloning experiments). Although Illumina NGS has not been considered suitable for QSR owing to its short read length, we for the first time successfully applied the QSR technique to NGS reads from MiSeq 2 x 300 nt paired-end sequencing. It was preliminarily confirmed that the SNV-to-SNV intervals were sufficiently short compared with the distribution of NGS reads obtained (Fig. 2B and S1 Fig.). Quasispecies sequences were successfully reconstructed in this study with a sufficient length covering the

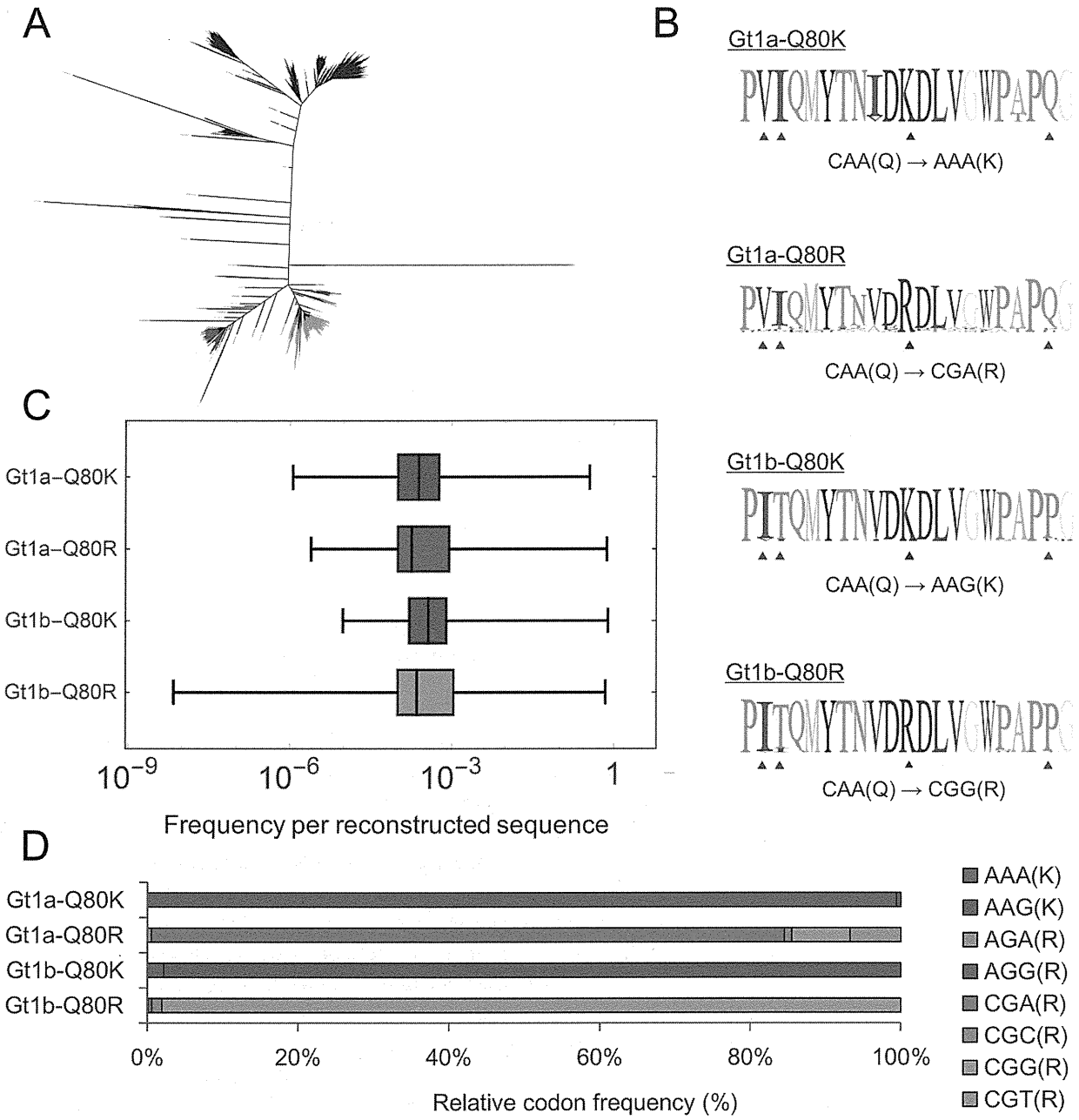

Fig 8. Integrated interdependence analysis of Gt and RAV enables high-throughput identification of distinct subpopulations with characteristic combinations of Gt and amino acid haplotype. (A) Phylogenetic analysis of all reconstructed sequences of NS3 protease region. Taxa are colored on the basis of their assigned Gt and Q80 RAV as follows: Gt1a-Q80K = blue; Gt1a-Q80R = purple; Gt1b-Q80K = red; Gt1b-Q80R = orange. A phylogenetic tree was generated using FigTree software. (B) Sequence logos from all sequences assigned to each pair of Gt and Q80 amino acid variant depicted in (A). Blue triangles denote NS3 Q80. Red triangles denote geno/subtype-specific amino acid polymorphisms at positions 71, 72, and 89. The codon change from reference to the most dominant variant at position 80 was denoted. (C) Distributions of estimated frequency per reconstructed sequence. (D) Relative codon frequencies for each Gt-RAV. The frequency was defined as the ratio of the number of reconstructed sequences possessing each codon and the total number of reconstructed sequences.

doi:10.1371/journal.pone.0119145.g008

core region and NS3 protease region. Furthermore, we achieved highly accurate estimations of Gts and RAVs by combining two QSR programs, QuRe and QuasiRecomb, both of which were based on different algorithm principles. Initially, we had a concern about artificial

recombination attributed to PCR amplification step and/or QSR calculation step. However, the simulation experiments demonstrated the accuracy of our QSR-based genotyping (Fig. 3 and Table 2) and RAV screening (Fig. 4 and Table 3) without *in silico* RAV recombination proven (Fig. 4A). A high PPV, rather than a high Sn, would be preferable for future investigation, because a high PPV would allow effective selection of patients having "true-positive" low-frequency RAVs, without the annoying false-positive RAVs. This is particularly important for research focusing on the impact of pre-existing minor RAVs, because a considerable number of false-positive RAVs at the preliminary screening stage might lead to a false conclusion that minor RAVs showed no correlation with the treatment outcome. In addition, note that sensitivity is in principle restricted by the coverage depth attainable with the sequencers currently available; therefore, so methodological improvement would be difficult. Lastly, by desterilizing the reconstructed haplotype information, we combined genotyping and RAV screening so as to find a novel relationship between them. The limitations of conventional SNV-based mutation screening are summarized into the following points: (1) it was difficult to gain genotype information; (2) it was difficult to link detected SNVs to correctly infer relevant RAVs, especially when multi-geno/subtype clones co-existed; and (3) it was impossible to gain insight on the basis of epistatic interactions between mutations, which has recently been predicted in HIV protease by a systems approach [45]. Our approach can overcome these limitations, which can reveal how the impact of one mutation depends on the presence or absence of other mutations in the context of clinical trials and post-trial surveys.

Recently, Jabara et al. have reported a novel solution to eliminating errors introduced during PCR amplification and pyrosequencing by using a single-molecular identifier [46,47]. The principle of this strategy is the use of a RT primer tagged by an 8 degenerate ID sequence. The resultant pyrosequencing reads having the same ID tag sequence are clustered, and the consensus sequence is generated on the majority basis, thus enabling the effective removal of artificial errors introduced during PCR, library preparation, and NGS. Polymerase error rate has been vigorously studied because of its potential impact on the inference of viral quasispecies diversity [48]. Although a promising technique, however, the analysis of this diversity could not yet be considered error-free owing to the error-prone nature of reverse transcriptase. The error rate of the commonly used, MMLV RTase was reported to be around 10^{-5}–10^{-4} per nucleotide [49], which might still be sufficient to artificially generate low-frequency false positive variants. Moreover, the read length of the NGS sequencer, a maximum of $\sim$1000 bp achieved using the Roche GS FLX+ system, would be an inevitable limitation of this methodology. Another solution that has recently been described by Acevedo et al. is circular sequencing (CirSeq), wherein circularized genomic RNA fragments are used to generate tandem repeats [50,51]. These repeated reverse transcriptions principally eliminate even the errors introduced by the reverse transcriptase use. The CirSeq approach in principle would provide completely error-free sequencing, but the target RNA must be fragmented into small pieces before amplification, which would be unfavorable for linkage analysis. In contrast to these emerging techniques, our analysis pipeline is much more practical. Moreover, our framework can be applicable even to previous NGS data obtained from ordinary RT-PCR experiments, as long as the read lengths are sufficiently large. NGS sequence meta-analysis is an emerging but promising strategy to integrate our knowledge leading to deeper insights on viral quasispecies dynamics.

Among hemophiliacs frequently receiving coagulation factor concentrates, the prevalence of HCV infection was high (60–90%) [52,53]. Before 1984, preheating was not yet routinely performed during the preparation of coagulation factor concentrates to inactivate contaminating HIV [54,55]. Moreover, blood products were frequently imported from countries overseas including the United States, as there were insufficient blood donors in Japan. Thus, patients using blood products at that time were exposed to the risk of infection with not only HCV but

also HIV, which had not yet been prevalent in Japan. Considering this specific circumstance, we hypothesized that there were HCV quasispecies of different genetic and geographic origins among HCV monoinfected non-hemophiliacs and HCV/HIV coinfected hemophiliacs in Japan. As expected, our analyses demonstrated that the compositions of genotypes and RAVs were quite different between HCV/HIV coinfected hemophiliacs, HCV monoinfected patients with previous exposure to whole-blood transfusion (BLx), and HCV monoinfected patients without a history of exposure to BLx. Gt1b was dominant (10 out of 11 = 91%) among cases without HIV coinfection, whereas Gt1a was dominant (6 out of 11 = 54%) among HCV/HIV coinfected patients. The predominant infection with Gt2a and Gt2b was determined in 3 cases. No other genotypes such as Gt3 and Gt4 were detected in this study. Moreover, multi-geno/subtype overlapping infection was significantly more prevalent among hemophiliacs and patients with BLx. This high prevalence of overlapping infection might explain the changes in genotype frequently observed among hemophiliacs [56,57] and other at-risk populations [58]. Furthermore, investigation on the interrelationships between Gts and RAVs suggests that Q80K was more prevalent in HCV/HIV coinfected hemophiliacs, whereas Q80R was less prevalent in HCV monoinfected non-hemophiliacs (Tables 4 and 5). A notable finding is that Q80K was significantly associated with Gt1b quasispecies among the hemophiliacs in this study (Tables 4 and 5, and Fig. 8). The Q80K variant is observed in 5.7–38% and 0.0–0.8% of patients with Gt1a and Gt1b HCV infections, respectively [42]. Q80K confers a 9.3-fold resistance against simeprevir in the Gt1a replicon system [42], and one clinical Phase 2 trial of simeprevir showed reduced SVR 24 rates with patient with Q80K mutation compared to those without Q80K (70.6–85.5% vs 55.0–66.7%) [38,59]. Currently, however, there is still limited information available regarding the impact of Q80K on Gt1b HCV infection, despite the fact that the effect of Q80K has been well characterized for Gt1a. In the first place, the epidemiology and characteristics of Gt1 sequences having the Q80K/R variant should be further studied, as searches of the Los Alamos database yielded a very unsatisfying number of previously identified sequences (S3 and S4 Tables). Detailed examination of the linkage between genotype and several RAVs may provide additional insights the clinical relevance of low-frequency genotype, drug-resistant quasispecies and their impact on the DAA therapy outcome.

Similarly to all studies, this study has some limitations. Firstly, the number of cases studied was very small, thus, the statistical power was insufficient to certainly detect low-prevalence mutations (e.g., R155, A156, and D168) if present. Secondly, although randomly selected, there might be a certain bias in enrolling HCV mono-infecting samples with a history of blood transfusion available. In this study, sample information including age, sex, associated illness, and source of infection, was not taken into consideration in the analysis; thus, the possibility of confounding and selection biases still remains. Thirdly, this study does not include hemophiliacs with HCV infection without HIV coinfection, because of sample unavailability. Finally, since this is a single-time point observational study, no information on the dynamic evolution of viral quasispecies is available. We are currently in process of another study targeting NS3 and NS5A using paired serum samples of pre-therapy, post-therapy, and post-relapse for hemophiliacs previously treated with peg-IFN plus ribavirin. We will also conduct post-trial surveillance of DAAs including simeprevir and sofosbuvir, wherein the NS3 and NS5B would be the target regions.

In conclusion, we developed and validated novel genotyping and RAV screening pipelines for HCV using the emerging technologies of NGS and QSR, reinforcing their potentials for the deconvolution of low-frequency genotypes, RAVs, and their interrelationships. Our study clearly demonstrated how the compositions of pre-existing minor genotypes and RAVs are considerably different between hemophiliacs and nonhemophiliacs, and HCV monoinfected patients with or without a history of whole-blood transfusion. These results strongly warrant

further studies investigating the epidemiology and impacts of low-frequency variants on the clinical outcome of DAA therapies among hemophiliacs and other high-risk populations.

Supporting Information

S1 Fig. Pairwise SNV-to-SNV distance distributions of HCV core and NS3 protease region estimated from Los Alamos HCV reference sequences. (A, B) SNV-to-SNV distance distributions of all possible Gt pairs of (A) core and (B) NS3 protease region were estimated from aligned reference sequences obtained from Los Alamos HCV sequence database. (C-F) Intragenotype (C, D) and intrasubtype (E, F) SNV-to-SNV distance distributions of (C, E) core and (D, F) NS3, respectively. A white notch represents median, and a red bar represents mean in each box-whisker chart.
(TIF)

S2 Fig. Phylogenetic positions of reconstructed sequences assigned to false-positive genotypes. Normalized patristic distances from reference sequences of each Gt were averaged, and distances from Gt1b and Gt2a were plotted. Sequences assigned to Gt1b were depicted in blue; Gt2a in cyan; Gt1a in red; Gt2b in orange; Gt2k in purple.
(TIF)

S3 Fig. Quantitative accuracy of different QSR methods for detecting minor genotypes. Reconstructed abundances under different simulation conditions were paired with corresponding preset abundance values. The conditions tested were as follows: (A, C) QuRe-Low and QuRe-High, and (B, D) QuasiRecomb-Low and QuasiRecomb-High, wherein Low represents the total read count of 30,000, and High represents 100,000 for (A, B) core and (C, D) NS3 protease region (NS3). Note that the abundance threshold was set to 0.001, and values below 0.001 were replaced with 0.001 for descriptive purposes.
(TIF)

S4 Fig. NS3 PI RAVs reproducibly detected by QuRe. QSR was performed using QuRe, and relative abundances of resistance-associated variants (RAVs) in the NS3 protease region were estimated in each subject. The x-axis labels are sample IDs colored on the basis of their history of exposure to blood (see Table 1 for details). The y-axis RAV labels were colored on the basis of the effects of RAVs on simeprevir susceptibility: susceptible (FC < 2) substitutions are in cyan, moderately resistant substitutions ($2 <$ FC < 50) in magenta. No highly resistant substitutions (FC > 50) were detected. The threshold was set at a frequency of 0.0001.
(TIF)

S5 Fig. NS3 PI RAVs reproducibly detected by QuasiRecomb. QSR was performed using QuasiRecomb, and relative abundances of resistance-associated variants (RAV) in the NS3 protease region were estimated in each subject. The x-axis labels are sample IDs colored on the basis of their status of exposure to blood (see Table 1 for details). The y-axis RAV labels are colored on the basis of the effects of RAVs on simeprevir susceptibility: susceptible (FC < 2) substitutions are in cyan, moderately resistant substitutions ($2 <$ FC < 50) in magenta; highly resistant substitutions (FC > 50) are in brown. The threshold was set at a frequency of 0.0001.
(TIF)

S1 Table. Primers used for RT-PCR.
(XLSX)

S2 Table. Parameter settings for QSR simulations.
(XLSX)

S3 Table. Counts of NS3 amino acid at position 80 binned by genotype in reported sequences retrieved from the Los Alamos HCV sequence database.
(XLSX)

S4 Table. Counts of NS3 amino acid at position 80 binned by sampled country in reported sequences retrieved from the Los Alamos HCV sequence database.
(XLSX)

Acknowledgments

We thank Kenji Chiba, Haruka Kukiya, Naoyuki Sugimoto, Tsukasa Okada, and Hiroshi Hashimoto (Hokkaido System Science Co. Ltd) for use of the MiSeq sequencing facility and technical assistance.

Author Contributions

Conceived and designed the experiments: MO HY. Performed the experiments: MO HY TT. Analyzed the data: MO HY TT HO WS. Contributed reagents/materials/analysis tools: HG SO. Wrote the paper: MO HY HO WS KM SK KK.

References

1. Messina JP, Humphreys I, Flaxman A, Brown A, Cooke GS, Pybus OG, et al. Global distribution and prevalence of hepatitis C virus genotypes. Hepatology. 2015; 61: 77–87. doi: 10.1002/hep.27259 PMID: 25069599
2. Scheel TKH, Rice CM. Understanding the hepatitis C virus life cycle paves the way for highly effective therapies. Nat Med. 2013; 19: 837–49. doi: 10.1038/nm.3248 PMID: 23836234
3. Asnis GM, De La Garza R. Interferon-induced depression in chronic hepatitis C: a review of its prevalence, risk factors, biology, and treatment approaches. J Clin Gastroenterol. 2006; 40: 322–335. PMID: 16633105
4. Chung RT, Baumert TF. Curing Chronic Hepatitis C—The Arc of a Medical Triumph. N Engl J Med. 2014; 370: 1–3. doi: 10.1056/NEJMp1313927 PMID: 24328440
5. Lee LY, Tong CYW, Wong T, Wilkinson M. New therapies for chronic hepatitis C infection: a systematic review of evidence from clinical trials. Int J Clin Pract. 2012; 66: 342–55. doi: 10.1111/j.1742-1241.2012.02895.x PMID: 22420497
6. Asselah T, Marcellin P. New direct-acting antivirals' combination for the treatment of chronic hepatitis C. Liver Int. 2011; 31 Suppl 1: 68–77. doi: 10.1111/j.1478-3231.2010.02411.x PMID: 21205141
7. Asselah T, Marcellin P. Second-wave IFN-based triple therapy for HCV genotype 1 infection: Simeprevir, faldaprevir and sofosbuvir. Liver Int. 2014; 34: 60–68. doi: 10.1111/liv.12424 PMID: 24373080
8. Schinazi R, Halfon P, Marcellin P, Asselah T. HCV direct-acting antiviral agents: The best interferon-free combinations. Liver Int. 2014; 34: 69–78. doi: 10.1111/liv.12423 PMID: 24373081
9. Everson GT, Sims KD, Rodriguez-Torres M, Hézode C, Lawitz E, Bourlière M, et al. Efficacy of an interferon- and ribavirin-free regimen of daclatasvir, asunaprevir, and BMS-791325 in treatment-naive patients with HCV genotype 1 infection. Gastroenterology. 2014; 146: 420–429. doi: 10.1053/j.gastro.2013.10.057 PMID: 24184132
10. Lawitz E, Sulkowski MS, Ghalib R, Rodriguez-Torres M, Younossi ZM, Corregidor A, et al. Simeprevir plus sofosbuvir, with or without ribavirin, to treat chronic infection with hepatitis C virus genotype 1 in non-responders to pegylated interferon and ribavirin and treatment-naive patients: the COSMOS randomised study. The Lancet. 26 Jul 2014: 1–10.
11. Eigen M. Viral quasispecies. Sci Am. 1993; 269: 42–9. PMID: 8337597
12. Ojosnegros S, Perales C, Mas A, Domingo E. Quasispecies as a matter of fact: viruses and beyond. Virus Res. 2011; 162: 203–15. doi: 10.1016/j.virusres.2011.09.018 PMID: 21945638
13. Kuntzen T, Timm J, Berical A, Lennon N, Berlin AM, Young SK, et al. Naturally occurring dominant resistance mutations to hepatitis C virus protease and polymerase inhibitors in treatment-naïve patients. Hepatology. 2008; 48: 1769–78. doi: 10.1002/hep.22549 PMID: 19026009

14. Paolucci S, Fiorina L, Mariani B, Gulminetti R, Novati S, Barbarini G, et al. Naturally occurring resistance mutations to inhibitors of HCV NS5A region and NS5B polymerase in DAA treatment-naive patients. Virol J. 2013; 10: 355. doi: 10.1186/1743-422X-10-355 PMID: 24341898

15. Metzner KJ, Giulieri SG, Knoepfel SA, Rauch P, Burgisser P, Yerly S, et al. Minority quasispecies of drug-resistant HIV-1 that lead to early therapy failure in treatment-naive and -adherent patients. Clin Infect Dis. 2009; 48: 239–47. doi: 10.1086/595703 PMID: 19086910

16. Donaldson EF, Harrington PR, O'Rear JJ, Naeger LK. Clinical evidence and bioinformatics characterization of potential hepatitis C virus resistance pathways for sofosbuvir. Hepatology. 2015; 61: 56–65. doi: 10.1002/hep.27375 PMID: 25123381

17. Abdelrahman T, Hughes J, Main J, McLauchlan J, Thursz M, Thomson E. Next-generation sequencing sheds light on the natural history of hepatitis C infection in patients who fail treatment. Hepatology. 2015; 61: 88–97. doi: 10.1002/hep.27192 PMID: 24797101

18. Abe H, Hayes CN, Hiraga N, Imamura M, Tsuge M, Miki D, et al. A translational study of resistance emergence using sequential direct-acting antiviral agents for hepatitis C using ultra-deep sequencing. Am J Gastroenterol. 2013; 108: 1464–72. doi: 10.1038/ajg.2013.205 PMID: 23896953

19. Loman NJ, Misra R V, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol. 2012; 30: 434–9. doi: 10.1038/nbt.2198 PMID: 22522955

20. Beerenwinkel N, Günthard HF, Roth V, Metzner KJ. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front Microbiol. 2012; 3: 329. doi: 10.3389/fmicb.2012.00329 PMID: 22973268

21. Zagordi O, Däumer M, Beisel C, Beerenwinkel N. Read length versus depth of coverage for viral quasispecies reconstruction. PLoS One. 2012; 7: e47046. doi: 10.1371/journal.pone.0047046 PMID: 23056573

22. Prosperi MCF, Yin L, Nolan DJ, Lowe AD, Goodenow MM, Salemi M. Empirical validation of viral quasispecies assembly algorithms: state-of-the-art and challenges. Sci Rep. 2013; 3: 2837. doi: 10.1038/srep02837 PMID: 24089188

23. Schirmer M, Sloan WT, Quince C. Benchmarking of viral haplotype reconstruction programmes: an overview of the capacities and limitations of currently available programmes. Brief Bioinform. 2014; 15: 431–42. doi: 10.1093/bib/bbs081 PMID: 23257116

24. Giallonardo F Di, Töpfer A, Rey M,Prabhakaran S,Duport Y,Leemann C, et al. Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. Nucleic Acids Res. 2014; 42: 1–12. doi: 10.1093/nar/gkt1324 PMID: 24376271

25. Prosperi MCF, Salemi M. QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. Bioinformatics. 2012; 28: 132–3. doi: 10.1093/bioinformatics/btr627 PMID: 22088846

26. Töpfer A, Zagordi O, Prabhakaran S, Roth V, Halperin E, Beerenwinkel N. Probabilistic inference of viral quasispecies subject to recombination. J Comput Biol. 2013; 20: 113–23. doi: 10.1089/cmb.2012.0232 PMID: 23383997

27. R Core Team. R: A language and environment for statistical computing. 2014.

28. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004; 5: R80. PMID: 15461798

29. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011; 17: 10.

30. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25: 1754–60. doi: 10.1093/bioinformatics/btp324 PMID: 19451168

31. Kuiken C, Yusim K, Boykin L, Richardson R. The Los Alamos hepatitis C sequence database. Bioinformatics. 2005; 21: 379–84. PMID: 15377502

32. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002; 30: 3059–66. PMID: 12136088

33. Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics. 2011; 12: 119. doi: 10.1186/1471-2105-12-119 PMID: 21521499

34. Macalalad AR, Zody MC, Charlebois P, Lennon NJ, Newman RM, Malboeuf CM, et al. Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. PLoS Comput Biol. 2012; 8: e1002417. doi: 10.1371/journal.pcbi.1002417 PMID: 22438797

35. Yang X, Charlebois P, Macalalad A, Henn MR, Zody MC. V-Phaser 2: variant inference for viral populations. BMC Genomics. 2013; 14: 674. doi: 10.1186/1471-2164-14-674 PMID: 24088188