Sequencing-based counting and size profiling of plasma Epstein–Barr virus DNA enhance population screening of nasopharyngeal carcinoma

W. K. Jacky Lam[a,b,c,d,1], Peiyong Jiang[a,b,c,1], K. C. Allen Chan[a,b,c,1], Suk H. Cheng[a,b], Haiqiang Zhang[a,b], Wenlei Peng[a,b], O. Y. Olivia Tse[a,b], Yu K. Tong[a,b], Wanxia Gai[a,b], Benny C. Y. Zee[e], Brigette B. Y. Ma[c,f], Edwin P. Hui[c,f], Anthony T. C. Chan[c,f], John K. S. Woo[d], Rossa W. K. Chiu[a,b,c], and Y. M. Dennis Lo[a,b,c,2]

[a]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong; [b]Department of Chemical Pathology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong; [c]State Key Laboratory in Oncology in South China, Sir Y. K. Pao Centre for Cancer, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong; [d]Department of Otorhinolaryngology, Head and Neck Surgery, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong; [e]Jockey Club School of Public Health and Primary Care, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong; and [f]Department of Clinical Oncology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong

Contributed by Y. M. Dennis Lo, April 26, 2018 (sent for review March 13, 2018; reviewed by Luis Diaz and Klaus Pantel)

Circulating tumor-derived DNA testing for cancer screening has recently been demonstrated in a prospective study on identification of nasopharyngeal carcinoma (NPC) among 20,174 asymptomatic individuals. Plasma EBV DNA, a marker for NPC, was detected using real-time PCR. While plasma EBV DNA was persistently detectable in 97.1% of the NPCs identified, ∼5% of the general population had transiently detectable plasma EBV DNA. We hypothesized that EBV DNA in plasma of subjects with or without NPC may have different molecular characteristics. We performed target-capture sequencing of plasma EBV DNA and identified differences in the abundance and size profiles of EBV DNA molecules within plasma of NPC and non-NPC subjects. NPC patients had significantly higher amounts of plasma EBV DNA, which showed longer fragment lengths. Cutoff values were established from an exploratory dataset and tested in a validation sample set. Adopting an algorithm that required a sample to concurrently pass cutoffs for EBV DNA counting and size measurements, NPCs were detected at a positive predictive value (PPV) of 19.6%. This represented superior performance compared with the PPV of 11.0% in the prospective screening study, which required participants with an initially detectable plasma EBV DNA result to be retested within 4 weeks. The observed differences in the molecular nature of EBV DNA molecules in plasma of subjects with or without NPC were successfully translated into a sequencing-based test that had a high PPV for NPC screening and achievable through single time-point testing.

liquid biopsy | massively parallel sequencing | circulating tumor DNA | ctDNA | size-based diagnostics

Liquid biopsies via circulating cell-free DNA analysis have been shown to be of value in noninvasive monitoring of cancer treatment response (1–3) and for the detection of cancer recurrence (4–6). To extend the application of circulating cell-free DNA to cancer screening, researchers have to face the challenge of developing assays that are sufficiently sensitive for detecting the expectedly low concentrations of circulating tumor DNA in early stages of cancer. Recently, our group demonstrated the feasibility of utilizing cell-free DNA for identifying early cancers through the detection of plasma Epstein–Barr virus (EBV) DNA for screening of nasopharyngeal carcinoma (NPC) among asymptomatic individuals (7). Circulating cell-free EBV DNA is a blood-based biomarker for EBV-related malignancies (8–10). Its clinical utility in prognostication and surveillance of recurrence of NPC has been validated (11, 12). To demonstrate the use of plasma EBV DNA for NPC screening, we conducted a large-scale prospective screening study that involved 20,174 asymptomatic participants identified from the community. A significantly higher proportion of early stage NPC cases (stage I or II) were identified in the screened cohort than in a historical unscreened cohort. The NPC cases identified by screening had longer progression-free survival. These promising results, together with the noninvasive nature of a blood-based test, would potentially contribute to widespread use of plasma EBV DNA as a screening tool for NPC.

Significance

We identified differentiating molecular characteristics of plasma EBV DNA between nasopharyngeal carcinoma (NPC) patients and non-NPC subjects. Sequencing-based analysis revealed higher amounts of plasma EBV DNA and generally longer fragment lengths of plasma viral molecules in NPC patients than in non-NPC subjects. Based on these findings, we have developed a highly accurate blood-based test for screening of NPC. Such an approach is shown to enhance the positive predictive value and demonstrate a superior performance for NPC screening. It also obviates the need of a follow-up blood sample and therefore allows single time-point testing. We believe that this more clinically practical protocol would facilitate NPC screening on a population scale.

Author contributions: W.K.J.L., P.J., K.C.A.C., R.W.K.C., and Y.M.D.L. designed research; W.K.J.L., S.H.C., Y.K.T., and W.G. performed research; W.K.J.L., S.H.C., Y.K.T., and W.G. performed the benchwork and the sequencing; W.K.J.L., P.J., K.C.A.C., H.Z., W.P., O.Y.O.T., B.C.Y.Z., B.B.Y.M., E.P.H., A.T.C.C., J.K.S.W., R.W.K.C., and Y.M.D.L. analyzed data; P.J., H.Z., W.P., and O.Y.O.T. performed the bioinformatics analysis; W.K.J.L., K.C.A.C., B.C.Y.Z., B.B.Y.M., E.P.H., A.T.C.C., and J.K.S.W. analyzed the clinical data; and W.K.J.L., P.J., K.C.A.C., R.W.K.C., and Y.M.D.L. wrote the paper.

Reviewers: L.D., Memorial Sloan Kettering Cancer Center; and K.P., Universitätsklinikum Hamburg-Eppendorf.

Conflict of interest statement: Y.M.D.L. is a scientific cofounder and member of the scientific advisory board for Grail. Y.M.D.L., R.W.K.C., and K.C.A.C. hold equity in Grail and receive research funding from Grail/Cirina. P.J. is a consultant to Xcelom and Grail. W.K.J.L. is a consultant to Grail. E.P.H. receives fees for serving on an advisory board for Bristol-Myers Squibb and Merck Sharp & Dohme. Y.M.D.L., R.W.K.C., K.C.A.C., P.J., and W.K.J.L. have filed patent applications based on the data generated from this work.

Data deposition: The sequence data reported in this paper (for the subjects studied in this work who consented to data archiving) have been deposited in the European Genome–Phenome Archive (EGA), https://www.ebi.ac.uk/ega/, hosted by the European Bioinformatics Institute (accession no. EGAS00001002707).

[1]W.K.J.L., P.J., and K.C.A.C. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: loym@cuhk.edu.hk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1804184115/-/DCSupplemental.

Published online May 14, 2018.

The real-time PCR assay used in the prospective screening study was shown to be highly sensitive to the detection of plasma EBV DNA even from small tumors. However, there is considerable room for improvement with respect to the test specificity. The peak age-specific incidence of NPC in Hong Kong is 40 per 100,000 persons (13), but ~5% of healthy population have detectable levels of EBV DNA in plasma (8, 14). The screening study yielded a positive predictive value (PPV) of 3.1% when plasma EBV DNA assessment by the real-time PCR assay was performed on one occasion per participant (7). To improve the PPV, participants who had detectable plasma EBV DNA were asked to provide another blood sample within 4 wk for a second test by the same PCR assay. Subjects were defined as "screen-positive" if they had any detectable EBV DNA in the plasma on both occasions. A previous study showed that EBV DNA tended to be persistently detectable in the plasma of NPC subjects but appeared transiently in the plasma of non-NPC subjects (15). The two time-point testing approach indeed reduced the false-positive rate from 5.4 to 1.4% with a resultant PPV of 11.0%. While the PPV of testing on two occasions fared well in comparison with population screening modalities for other cancers (16), in this study we asked whether there might be strategies for further enhancing the PPV for NPC screening by plasma EBV DNA analysis.

Previously, we have demonstrated that circulating EBV DNA molecules from NPC patients are released from NPC cells (17, 18). The source of the plasma EBV DNA must be different in persons without NPC. Unlike the persistent release of EBV DNA into the circulation from NPC cells, the source of EBV DNA only contributes such DNA transiently in the persons without NPC. These observations led us to hypothesize that the molecular nature of the EBV DNA might be different in the plasma of persons with and without NPC. Different molecular characteristics have been identified among circulating DNA molecules originating from different human tissues (19–22). Studies on the size profile of plasma DNA (19, 20, 23–25) and its relationship to nucleosomal organization (19, 21, 26) have been actively pursued. We have reported the different size profiles of cell-free maternal DNA and fetal DNA in the maternal circulation (19, 27, 28). Circulating maternal DNA consists of a predominant population of plasma DNA molecules with a size of 166 bp, whereas fetal DNA consists of a predominant population with a size of 143 bp. Similarly, in the context of cancer, we showed that circulating tumor-derived DNA molecules were shorter than non–tumor-derived DNA in patients with hepatocellular carcinoma (20). Extending beyond studies on plasma DNA molecules of human origin, here we propose that there might be differences in the molecular characteristics between viral DNA molecules present in the plasma of persons with or without NPC. Differences among plasma viral DNA molecules, if present, would allow one to better differentiate NPC patients, thus reducing the false-positive rate and enhancing the PPV of a plasma EBV DNA-based screening test. These improvements might allow us to replace the two time-point testing protocol with one based on testing at a single time point.

Results

Study Design and Case Assignment. We used targeted sequencing to capture and measure the abundance as well as the size profiles of EBV DNA in plasma of subjects with and without NPC (Fig. 1). The aim of the study was to identify plasma EBV DNA molecular characteristics that would ideally maintain the sensitivity but would offer enhanced specificity in differentiating persons with NPC from the ~5% of the population who did not have NPC but had positive plasma EBV DNA (8, 14). Therefore, this study could only be conducted on sample sets where the person's clinical status had already been determined based on plasma EBV DNA real-time PCR assessment. The plasma samples that we had collected in our recently conducted prospective NPC screening trial (7) represented a unique resource for investigating NPC-specific molecular signatures of plasma EBV DNA. Fig. 2 shows the study protocol of the prospective screening study and the distribution of cases with positive plasma EBV detection when tested by real-time PCR.

The protocol of the previous screening study (7) involved two time-point testing by the same real-time PCR assay (Fig. 2*A*). Out of the recruited 20,174 asymptomatic Chinese males aged between 40 and 62 y, there were 1,112 subjects (5.5%) who had detectable plasma EBV DNA from baseline PCR tests (Fig. 3). Among them,

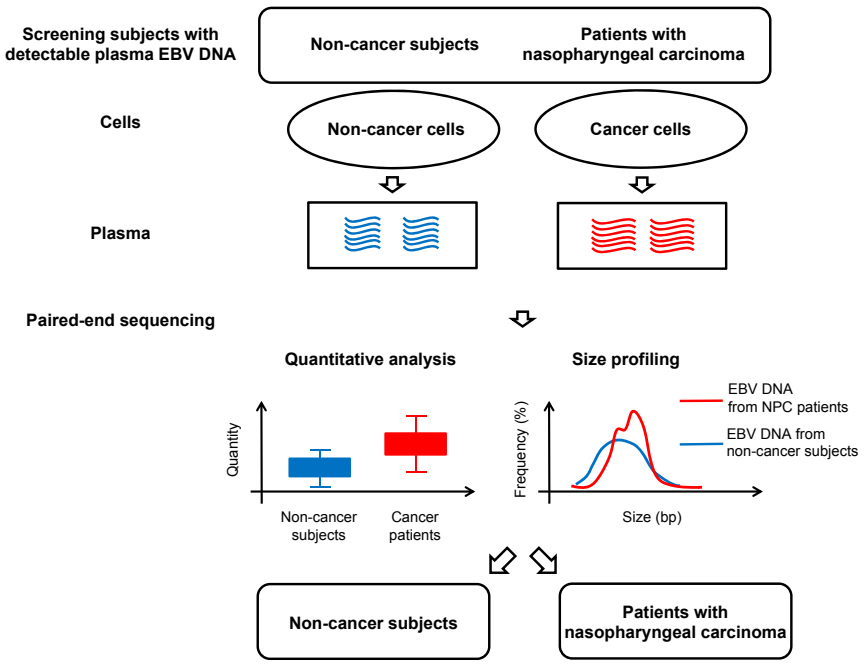

Fig. 1. Quantitative and size profile analysis of plasma EBV DNA in patients with nasopharyngeal carcinoma (NPC) and non-NPC subjects.

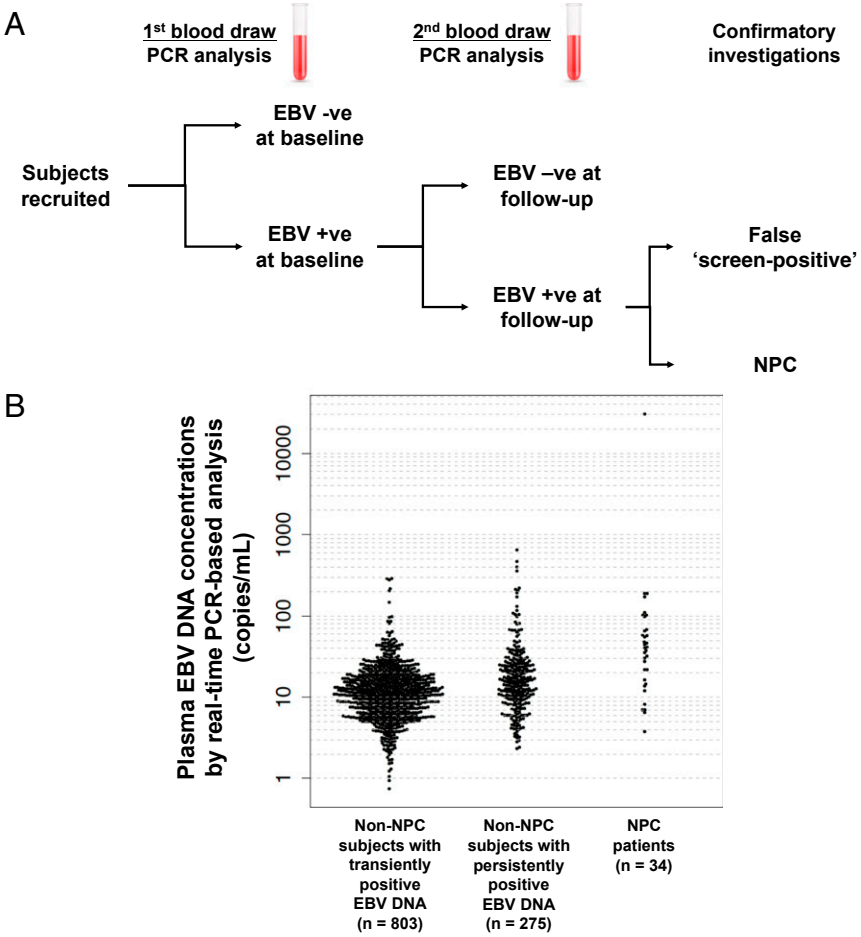

Fig. 2. Overview of the prospective screening study and subject classification based on real-time PCR. (*A*) The study protocol of the screening study. The protocol involved a two time-point testing of plasma EBV DNA by the same real-time PCR assay. Subjects who had detectable plasma EBV DNA at both baseline and follow-up tests were defined as screen-positive. Screen-positive subjects would be referred for confirmatory investigations. (*B*) Plasma EBV DNA concentrations by quantitative PCR-based analysis in non-NPC subjects with transiently positive and persistently positive plasma EBV DNA results and patients with NPC in the screening cohort are shown.

34 subjects were later confirmed to have NPC. For the remaining 1,078 non-NPC subjects, 803 subjects had "transiently positive" plasma EBV DNA results (i.e., positive at baseline but negative at follow-up) and 275 had "persistently positive" plasma EBV DNA results (i.e., positive at both baseline and follow-up). In the study (7), plasma EBV DNA results were expressed as "positive" or "negative." Here, we reviewed the levels of the plasma EBV DNA concentrations between the groups as measured by real-time PCR (Fig. 2*B*). The mean plasma EBV DNA concentration of the NPC group [942 copies per mL; interquartile range (IQR), 18–68 copies per mL] was significantly higher than those of the transiently positive group (16 copies per mL; IQR, 7–18 copies per mL) and persistently positive group (30 copies per mL; IQR, 9–26 copies per mL) ($P < 0.0001$, Kruskal–Wallis test). However, there is much overlap in the plasma EBV DNA concentrations among the three groups (Fig. 2*B*).

In the current study, we first explored whether differences in the molecular profiles of plasma EBV DNA existed between persons with positive plasma EBV DNA associated with or not associated with NPC. We randomly selected 10 NPC patients and 40 non-NPC subjects (20 with transiently positive and 20 with persistently positive EBV DNA results) from the screening study to be included in the exploratory sample set (Fig. 3). Among the 10 randomly selected NPC subjects, 5 had stage I, 2 had stage II, 2 had stage III, and 1 had stage IV diseases. The subject characteristics are shown in Table 1. In the validation sample set, we included the remaining 24 patients with NPC and randomly selected 232 non-NPC subjects from the screening study (Fig. 3). There is no statistically significant difference in the plasma EBV DNA concentrations measured by real-time PCR among the selected NPC patients in the exploratory set and those in the validation set from the screening cohort ($P = 0.2$, t test). These 232 non-NPC subjects included 159 subjects with transiently positive and 73 with persistently positive plasma EBV DNA results. The ratio of the transiently positive group to the persistently positive group is similar to the actual ratio observed in the screening study. There is no statistically significant difference in the plasma EBV DNA concentrations measured by real-time PCR between these selected 232 non-NPC subjects and all non-NPC subjects in the screening cohort ($P = 0.07$, t test). Subjects in the validation group did not overlap with those in the exploratory group. We have also included 31 other NPC patients from an external unscreened population in the validation sample set (Fig. 3). Among the 24 NPC patients from the screening cohort, 11 had stage I, 6 had stage II, 6 had stage III, and 1 had stage IV diseases. There is no statistically significant difference in the distribution of NPC patients with early stage (stages I and II) and late-stage (stages III and IV) diseases from the screening cohort between the exploratory and validation sample sets ($P = 1.0$, Fisher's exact test). Among the 31 NPC patients from the

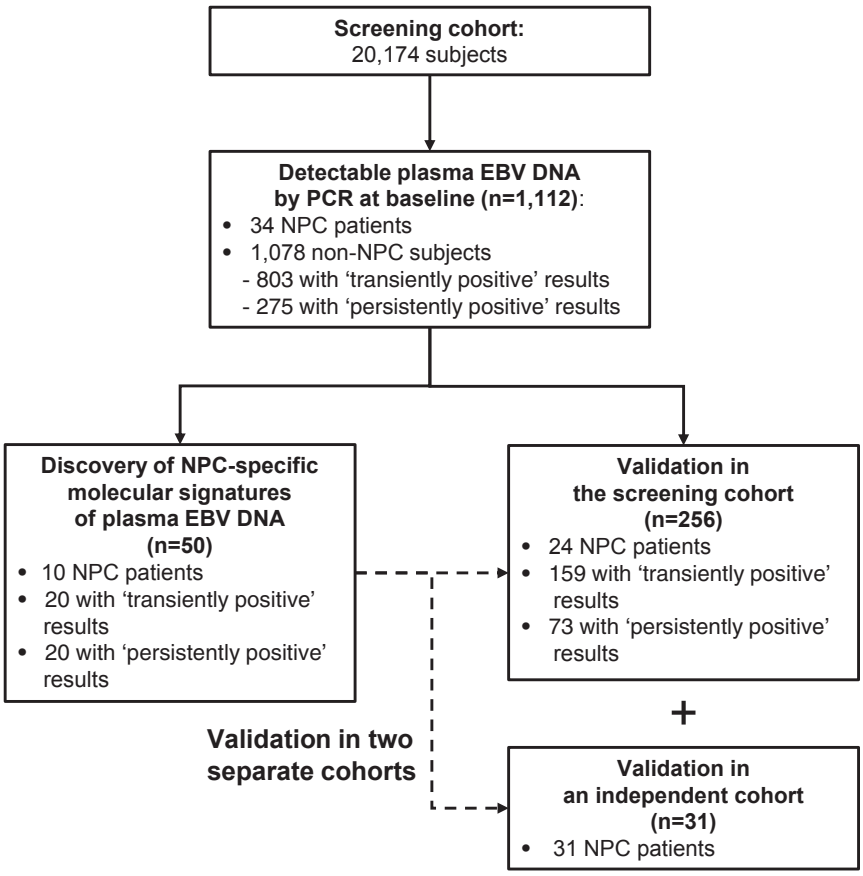

Fig. 3. Study cohorts. In the screening study, 20,174 subjects were recruited, and all received a baseline test for plasma EBV DNA by real-time PCR. In total, 1,112 subjects had detectable plasma EBV DNA at baseline. Among them, 34 subjects were confirmed to have NPC. For the remaining 1,078 non-NPC subjects, 803 subjects had transiently positive plasma EBV DNA results (i.e., positive at baseline but negative at follow-up) and 275 had persistently positive plasma EBV DNA results (i.e., positive at both baseline and follow-up). Plasma samples of NPC and non-NPC subjects were randomly selected and distributed into the exploratory and validation sample sets for the current study. All of the 34 NPC cases from the screening study had been analyzed either as part of the exploratory or validation sample sets. An additional 31 NPC patients from an independent cohort were included in the validation sample set.

external cohort, there were 3 patients with stage I, 2 with stage II, 20 with stage III, and 6 with stage IV diseases.

The sequencing analyses were performed on the baseline (first time-point) sample collected at the time of enrollment into the prospective screening study. For all of the samples in the exploratory and validation sample sets, the median number of mapped reads per sample was 70 million (IQR, 61 million to 85 million).

Sequencing-Based Quantification of Plasma EBV DNA in the Exploratory Sample Set. Plasma DNA molecules were captured by probes covering the entire EBV genome and portions of

Table 1. Subject characteristics in the exploratory and validation sample sets

	Exploratory dataset			Validation dataset			
Characteristics	Non-NPC subjects with transiently positive plasma EBV DNA	Non-NPC subjects with persistently positive plasma EBV DNA	NPC patients from the screening cohort	Non-NPC subjects with transiently positive plasma EBV DNA	Non-NPC subjects with persistently positive plasma EBV DNA	NPC patients from the screening cohort	NPC patients from an external cohort
Number	20	20	10	159	73	24	31
Sex							
M	20	20	10	159	73	24	24
F	0	0	0	0	0	0	7
Median age, y (IQR)	54.5 (50–56)	55 (50–60.5)	54 (47–56)	53 (47–57)	53 (48–59)	50 (43–54)	56 (50–62)
Tumor stage							
I			5			11	3
II			2			6	2
III			2			6	20
IV			1			1	6

human chromosomes 1, 2, 3, 5, 8, 15, and 22, and then sequenced. Plasma EBV DNA reads referred to plasma DNA fragments that were sequenced and mapped to the EBV genome. We measured the proportion of EBV DNA reads among the total number of sequenced DNA reads after removal of PCR duplicates. This would be subsequently referred to as the count-based analysis. Patients with NPC (median, 7.6×10^{-5}; IQR, 6.2×10^{-5} to 1.1×10^{-4}) had a statistically significantly higher proportion of EBV DNA reads than non-NPC subjects with transiently positive (median, 6.9×10^{-6}; IQR, 1.1×10^{-6} to 1.9×10^{-5}; $P = 0.0005$, Kruskal–Wallis test) and persistently positive results (median, 3.0×10^{-5}; IQR, 4.5×10^{-6} to 5.8×10^{-5}; $P = 0.04$, Kruskal–Wallis test) (Fig. 4*A*).

Size Profiling of Plasma EBV DNA by Sequencing in the Exploratory Sample Set. We studied and analyzed the differences in the size distributions of plasma EBV DNA reads from NPC patients and non-NPC subjects in the exploratory sample set. The size of each

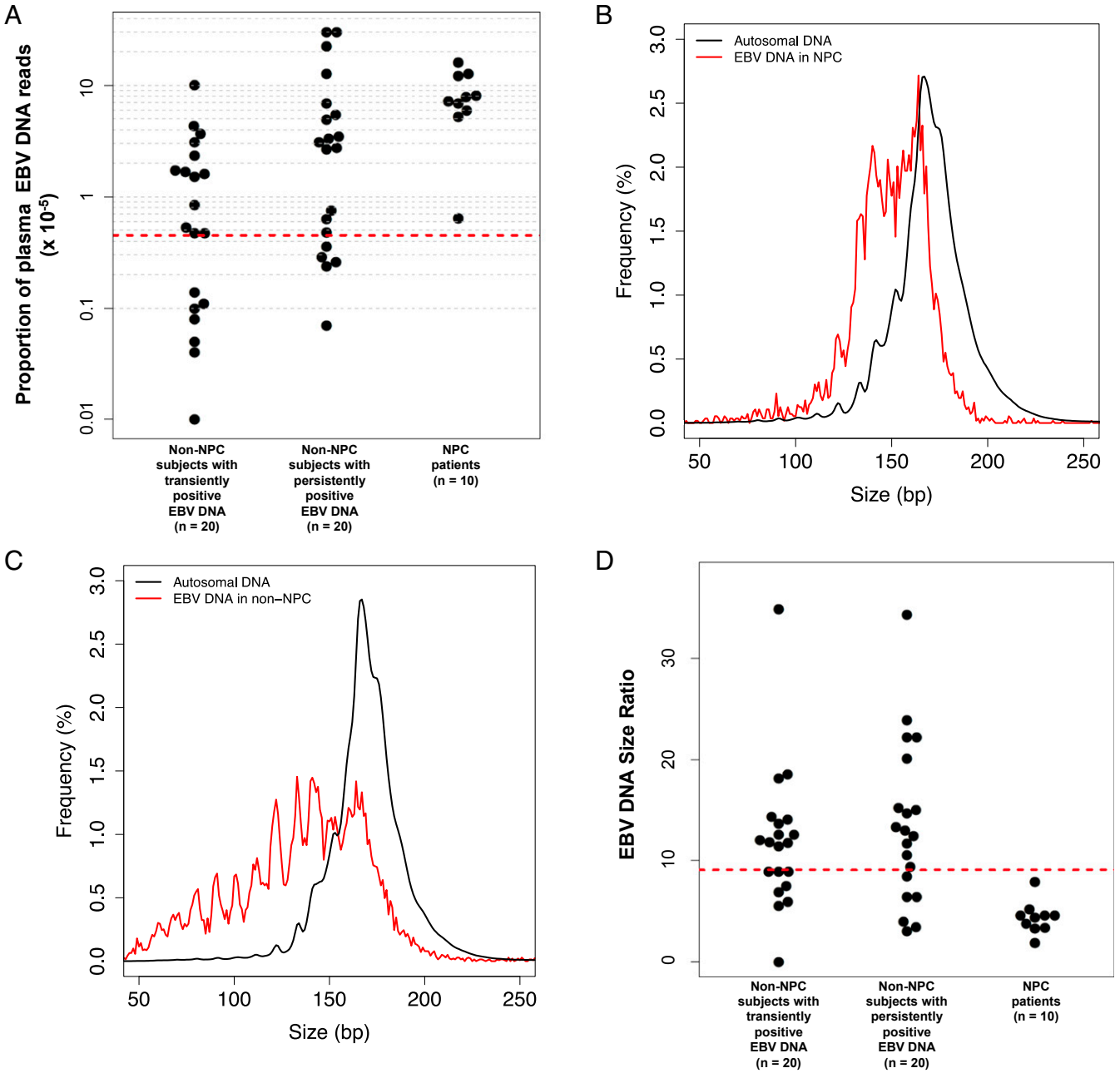

Fig. 4. Count-based and size-based analyses of plasma EBV DNA by target-capture sequencing in the exploratory sample set. (*A*) The proportions of plasma EBV DNA reads among the total number of sequenced plasma DNA reads of the NPC patients and non-NPC subjects with transiently positive and persistently positive results in the exploratory dataset are shown. A cutoff value was defined at 3 SDs below the mean of the logarithmic values of the proportion of EBV DNA reads of the 10 NPC patients in exploratory dataset. The cutoff value of 4.5×10^{-6} is denoted by the red dotted line. (*B*) Size distributions of EBV DNA (red curve) and autosomal DNA (black curve) in the plasma of a patient with NPC. (*C*) Size distributions of EBV DNA (red curve) and human autosomal DNA (black curve) in a non-NPC subject with persistently positive plasma EBV DNA results. (*D*) The EBV size ratios of NPC and non-NPC cases in the exploratory sample set are shown. A cutoff value was defined at 3 SDs above the mean values of the EBV size ratios of all of the 10 NPC patients in the exploratory dataset. The cutoff value of 9.1 is denoted by the red dotted line.

sequenced plasma DNA molecule was derived from the start and end coordinates of the paired-end reads. In Fig. 4 *B* and *C*, the size profiles of EBV DNA from a representative case of a patient with NPC and a non-NPC subject with persistently positive EBV DNA result are shown. We observed that the size profiles of EBV DNA from NPC patients exhibited a reduction in the 166-bp peak, but with a more pronounced peak at around 150 bp compared with the size profiles of human autosomal DNA (Fig. 4*B*). The size profiles of EBV DNA from non-NPC subjects showed peaks that were distributed over the shorter fragment sizes (Fig. 4*C*). Thus, NPC patients had a lower proportion of EBV DNA molecules shorter than 110 bp compared with that of the non-NPC subjects.

We developed a metric, the EBV DNA size ratio, to indicate the relative proportion of short EBV DNA fragments within each sample. The size ratio was defined as the proportion of short EBV DNA fragments with sizes within 80–110 bp normalized by that of autosomal DNA fragments within the same size range. Plasma EBV DNA fragments within the size range of 80–110 bp were deemed as short fragments because the resultant EBV DNA size ratio yielded the best discrimination between NPC patients and non-NPC subjects in the exploratory sample set compared with other size ranges (*SI Appendix*, Fig. S1). The lower the EBV DNA size ratio, the lower the proportion of EBV DNA molecules of sizes between 80 and 110 bp. It was calculated using the following equation:

$$\text{EBV DNA Size Ratio} = \frac{\text{Proportion of EBV DNA within } 80-110 \text{ bp}}{\text{Proportion of autosomal DNA within } 80-110 \text{ bp}}.$$

The median size ratio from samples of NPC patients (median, 4.5; IQR, 3.5–4.6) was significantly lower than the median ratios from samples of non-NPC subjects with transiently (median, 11.8; IQR, 8.6–13.8; $P = 0.001$, Kruskal–Wallis test) or persistently positive plasma EBV DNA (median, 12.7; IQR, 8.0–16.5; $P = 0.0005$, Kruskal–Wallis test) (Fig. 4*D*). There is no statistically significant difference in the median size ratios between non-NPC subjects with transiently and persistently positive plasma EBV DNA ($P = 0.5$, Mann–Whitney U test).

Defining the Cutoffs for the Count- and Size-Based Analyses. Using the data derived from the exploratory sample set, cutoff values in the count-based and size-based analyses were defined to achieve 100% sensitivity for capturing all of the NPC cases. In the count-based analysis, a cutoff value was defined at 3 SDs below the mean of the logarithmic values of portion of EBV DNA reads of these 10 NPC patients in the exploratory dataset. The cutoff value of 4.5×10^{-6} was obtained (Fig. 4*A*). Using this cutoff value, 13 of the 20 subjects with transiently positive and 15 of the 20 subjects with persistently positive EBV DNA results passed the cutoff in the count-based analysis.

Similarly, in the size-based analysis, a cutoff value was defined at 3 SDs above the mean values of the EBV DNA size ratios of all of the 10 patients. The cutoff value of 9.1 was obtained (Fig. 4*D*). Using this cutoff value, 8 of 20 subjects with transiently positive and 6 of 20 subjects with persistently positive EBV DNA results passed the cutoff in the size-based analysis.

Validation of Count- and Size-Based Analyses in a Validation Sample Set. We analyzed the proportions of EBV DNA reads in all of the samples from the validation sample set. There were significantly higher proportions of EBV DNA reads from samples of NPC patients from both the screening cohort (median, 2.2×10^{-4}; IQR, 8.9×10^{-5} to 1.5×10^{-3}) and the external cohort (median, 1.7×10^{-3}; IQR, 2.5×10^{-4} to 5.4×10^{-3}) than samples of non-NPC subjects with transiently (median, 2.1×10^{-6}; IQR, 6.5×10^{-7} to 8.0×10^{-6}; $P < 0.0001$) and persistently positive results (median, 2.4×10^{-5}; IQR, 1.1×10^{-5} to 5.0×10^{-5}; $P = 0.0044$). With the cutoff value of 4.5×10^{-6} defined in the exploratory dataset, all of the samples of NPC patients from both cohorts could be captured and had proportions of EBV DNA reads higher than the defined cutoff value. There were 56 (out of 159) subjects with transiently positive results and 64 (out of 73) subjects with persistently positive results who passed the cutoff in the count-based analysis (Fig. 5*A*).

In Fig. 5*B*, lower EBV DNA size ratios were observed in samples of NPC patients from both the screening (median, 3.2; IQR, 2.4–4.2) and external cohorts (median, 3.0; IQR, 2.4–4.3) than samples of non-NPC subjects with transiently positive (median, 11.3; IQR, 7.6–15.1; $P < 0.0001$) and persistently positive results (median, 12.7; IQR, 9.0–16.5; $P < 0.0001$). These results demonstrated that our finding of a lower proportion of short EBV DNA fragments in patients with NPC from the exploratory dataset was also observed in the external cohort. With the cutoff value of 9.1 defined in the exploratory dataset, all of the samples of NPC patients from both cohorts had EBV DNA size ratio smaller than the cutoff value. There were 55 (out of 159) subjects with transiently positive results and 19 (out of 73) subjects with persistently positive results who passed the cutoff in the size-based analysis.

Combined Count- and Size-Based Plasma EBV DNA Analysis in the Validation Sample Set. We assessed the value of combining the count- and size-based analyses for NPC identification. In this combined analysis, a plasma sample was deemed to be positive and classified as NPC if its sequencing data concurrently passed the cutoffs in both the count- and size-based analyses. We performed the combined count- and size-based analysis for all samples in the validation sample set. By applying the same cutoff values defined in the exploratory dataset, all of the samples of NPC patients from both the screening and external cohorts could be captured. There were 15 (out of 159) subjects with transiently positive results and 17 (out of 73) subjects with persistently positively results who passed both the cutoffs in the count- and size-based analyses (Fig. 5*C*). We compared the diagnostic performances of the count-based, size-based, and combined count- and size-based analyses as well as real-time PCR in differentiating NPC patients from non-NPC subjects in the validation sample set using receiver operating characteristic (ROC) curve analysis. Area under the curve values for the count-based, size-based, and combined analyses were 0.93, 0.92, and 0.97, respectively, and were significantly higher than that of PCR-based analysis (0.75) ($P = 0.0071$, $P = 0.0143$, $P = 0.0008$, respectively; bootstrap test) (Fig. 5*D*).

Modeling the Performance of Sequencing-Based Analysis of Plasma EBV DNA in the Screening Cohort. Since the combined count- and size-based analysis achieved the best diagnostic performance, we proposed an NPC screening protocol with incorporation of the combined analysis of plasma EBV DNA after baseline real-time PCR-based analysis (Fig. 6). Target-capture sequencing would be performed on baseline plasma samples with EBV DNA detectable by real-time PCR. In this protocol, subjects are defined as screen-positive if their plasma samples pass the cutoffs in both the count-based and the size-based analyses. Subjects are defined as screen-negative if the plasma samples do not pass the cutoff in either the count-based or size-based analysis. Based on the performance of the combined analysis in the validation sample set, we have estimated the sensitivity, specificity, positive predictive value, and false-positive rate in the entire 20,174-subject cohort of the prospective screening trial assuming the screening protocol were adopted. In the validation cohort, 15 out of 159 (9.4%) subjects with transiently positive EBV DNA results, 17 out of 73 (23.3%) subjects with persistently positive results,

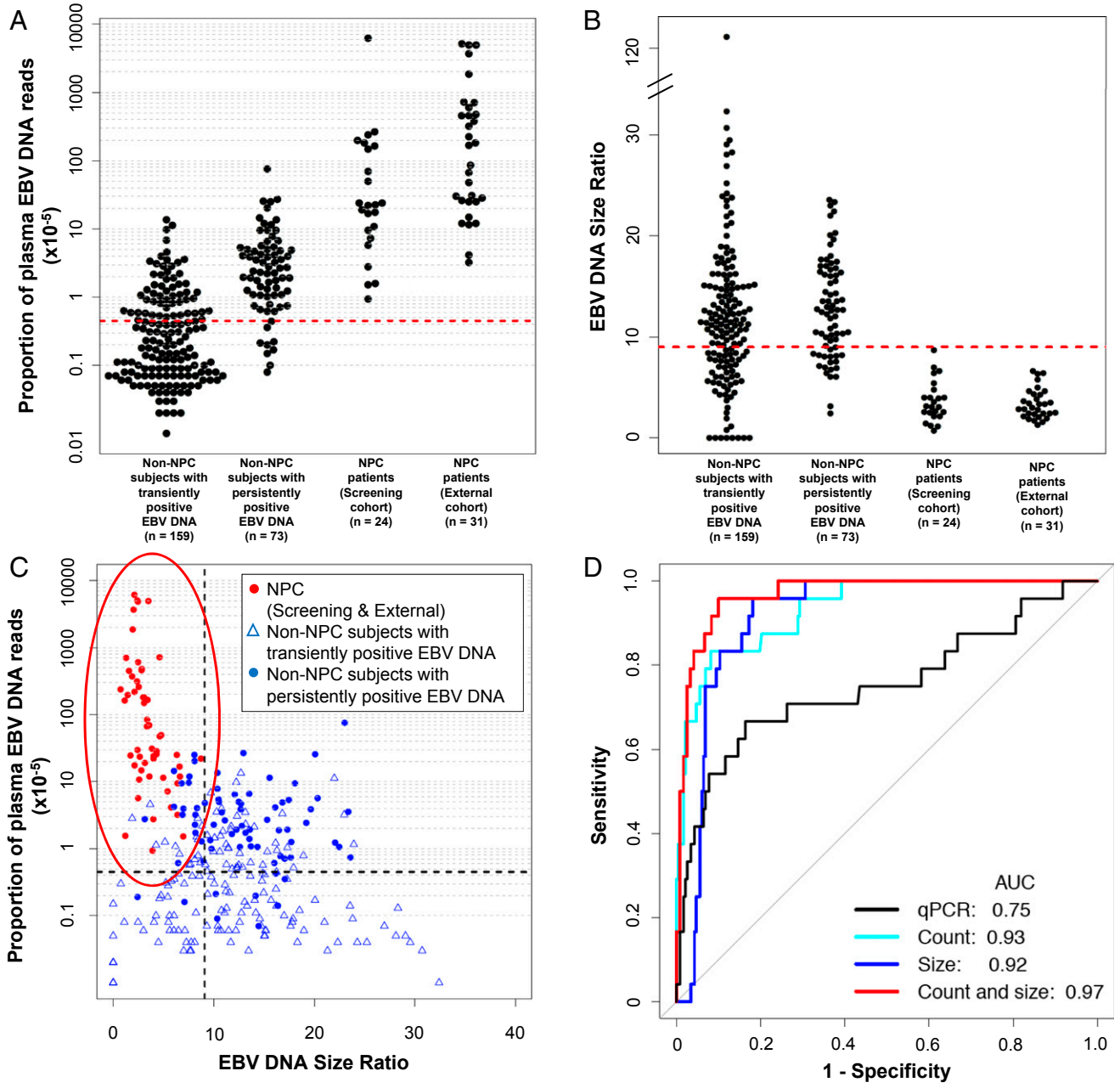

Fig. 5. Count- and size-based analyses of plasma EBV DNA by target-capture sequencing in the validation sample set. (*A*) The proportions of plasma EBV DNA reads of the NPC patients and non-NPC subjects are shown. The same cutoff value of 4.5×10^{-6} defined in exploratory dataset is denoted by the red dotted line. (*B*) The EBV DNA size ratios of the NPC patients and non-NPC subjects are shown. The same cutoff value of 9.1 defined in the exploratory dataset is denoted by the red dotted line. (*C*) Plot of the proportions of plasma EBV reads and corresponding size ratio values for all of the cases in the validation sample set. The same cutoff values in the count- and size-based analyses defined in the exploratory sample set are denoted by the gray dotted lines. The red oval highlights the quadrant with cases that passed the cutoffs in the combined count- and size-based analysis. (*D*) ROC curves for the count-based analysis, size-based analysis, combined sequencing analysis, and real-time PCR analysis are shown. Area under the curve (AUC) values are shown.

and all of the 24 patients with NPC (100%) passed the cutoffs in the count- and size-based analyses. These subjects were all considered as screen-positive according to the protocol. In the screening study, one subject with undetectable plasma EBV DNA by real-time PCR analysis developed NPC within 1 y (7). Since all of the NPC cases tested positive in the screening study could be captured, the projected sensitivity of this protocol would be 97.1% (95% CI, 85.1–99.9%), which would be the same as for the previous two time-point screening protocol. The estimated number of subjects with false screen-positive results was 140 (9.4% of the transiently positive group and 23.3% of the persistently group in the screening cohort). The estimated specificity would be 99.3% (95% CI, 99.2–99.4%). The PPV and false-positive rate were estimated to be 19.5% (95% CI, 13.9–26.2%) and 0.70% (95% CI, 0.59–0.82%), respectively (Fig. 6). The projected performance of the count-based and size-based analyses in the 20,174-subject cohort is shown in Table 2.

Discussion

In the current study, we have demonstrated that the molecular profiles of plasma EBV DNA determined by target-capture sequencing are different between persons with or without NPC.

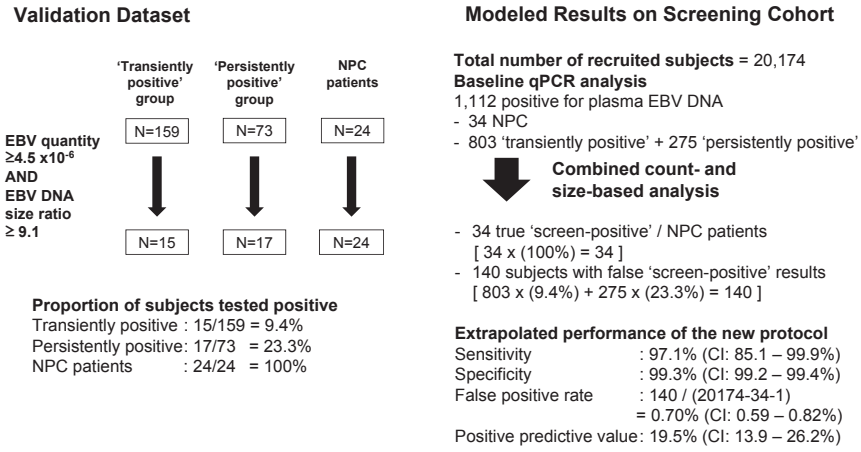

Fig. 6. Modeling the performance of sequencing-based analysis of plasma EBV DNA in the entire 20,174-subject screening cohort. The estimated sensitivity, specificity, positive predictive value, and false-positive rates are stated. CI denotes 95% CIs.

This observation raises interesting questions regarding the biological origins of the EBV DNA molecules in the plasma of individuals tested positive but without evidence of NPC. Previously, we used multiple PCR assays of different amplicon sizes to study the size distributions of plasma EBV DNA molecules from patients with NPC (17). We found that the circulating EBV DNA molecules from NPC patients were predominantly shorter than 181 bp. Those observations led us to conclude that the EBV DNA molecules were not associated with viral particles but instead were released from the cancer cells as part of the process of cell death. Paired-end massively parallel sequencing provides much more precise DNA size measurement whereby the size of each plasma DNA molecule could be determined. This is unlike PCR where the amount of DNA molecules at least as long as the intended amplicon is measured. In this study, we observed that the size distributions of EBV DNA from NPC patients (Fig. 4*B*) exhibited a reduction in the 166-bp peak and a relative prominence at around 150 bp compared with plasma DNA of human origin. The presence of the characteristic 166-bp peak in the plasma EBV DNA size profile of NPC patients suggested that circulating EBV DNA was nucleosome-bound. The relative prominence of EBV DNA (as circulating tumor DNA) at around 150 bp was concordant with our previous finding that tumor-derived DNA was in general shorter than non–tumor-derived DNA (20). In contrast to NPC patients, plasma EBV DNA from non-NPC subjects did not exhibit the typical nucleosomal pattern (Fig. 4*C*). Virion-associated EBV DNA has been reported to be free of nucleosomes (29, 30). We postulate that the lack of protection by nucleosomes (30) may render degradation products from virion-derived DNA having a shorter size distribution than plasma EBV DNA from NPC patients.

The differences in size profiles of EBV DNA molecules in plasma of NPC patients and that of non-NPC subjects may also contribute to the different quantitative profiles of plasma EBV DNA when assessed by real-time PCR and that by sequencing.

Table 2. Projected diagnostic performance of the count-based, size-based, and combined analyses in the 20,174-subject cohort

	20,174-Subject cohort		
Analysis	Sensitivity, %	Specificity, %	PPV, %
Count-based analysis	97.1	97.4	6.1
Size-based analysis	97.1	98.3	8.9
Combined analysis	97.1	99.3	19.6

PCR could only quantify EBV DNA molecules where the amplicon region is intact. EBV DNA molecules that are fragmented at a location within the amplicon cannot be amplified by the PCR assay. Sequencing, on the other hand, is not subjected to such a restriction and any EBV DNA fragments can be analyzed and counted.

By uncovering the differences in abundance and size profiles of plasma EBV DNA among subjects with or without NPC, more specific identification of NPC could be achieved. The cutoff values determined for the sequencing analyses were based on the goal to maintain the sensitivity achieved with real-time PCR testing in the prospective screening study (7). Thus, the key differentiator for the performance of the sequencing-based test was rooted in the improved specificity that it offered. The data from the prospective screening study showed that real-time PCR had a false-positive rate of 5.3% when testing was performed on a single-baseline blood sample. Such a test was associated with a PPV of 3.1%. In this study, the sequencing analyses were performed on the baseline blood samples. Using the count-based sequencing approach, the false-positive rate was 2.6%, and the PPV was 6.0% (Table 2). The size-based approach when used independently had a false-positive rate of 1.7% and PPV of 8.9%. To further enhance the test performance, we proposed an approach that required a sample to concurrently pass both the count- and size-based cutoffs to be deemed as tested positive. Using such a combined approach, we achieved a false-positive rate of just 0.7% and PPV of 19.6%. These data represent a significant improvement over that of real-time PCR even when considering the two time-point protocol that had a PPV of 11.0%.

From the public health point of view, any improvement in the PPV would have a substantial impact. Guangdong is one of the provinces in China with the highest incidence rates of NPC (31). According to the China Statistical Yearbook 2016, there are about 20 million men aged between 40 and 65 y in the Guangdong Province. If a universal NPC screening program is adopted for all men within this age range who have the highest age-specific incidence, the currently proposed protocol of combined analysis would lead to a reduction in the number of false positives by 50%, that is, 140,000 subjects. Such a large number would imply a substantial reduction in the medical expenditures initially spent on follow-up tests and confirmatory investigations including endoscopy and magnetic resonance imaging (MRI).

One key additional benefit of our sequencing-based approach is that high PPV could be achieved from blood sampling of just a single time point. By overcoming the need for testing at two time

points, our protocol has significant advantages over the previous protocol requiring two time-point testing. Previous studies on other cancer screening programs have shown that a substantial proportion of participants with abnormal screening test results reported anxiety and distress (32). Hence, for two time-point testing, subjects with detectable EBV DNA at baseline could only obtain their final screening status after the follow-up tests. These subjects may experience anxiety while waiting for the follow-up tests. In addition, the requirement to test two time points presents a number of logistical challenges. There are direct costs associated with recalling subjects initially tested positive for second testing. Compliance may become an issue when the two time-point protocol is adopted in the clinical context. Reduction in compliance would result in a reduction in the sensitivity of the NPC screening program. In contrast, the sequencing-based protocol obviates the need for a second blood sample to define a person having been "screened positive." This protocol is thus more clinically and logistically practical, and may be more easily accepted by the population to be screened.

The data in our current study showed that target sequencing analysis of plasma EBV DNA could achieve high PPV for NPC screening. In this study, the performance of the sequencing test is modeled against the cases assessed by real-time PCR in the prospective screening study. Therefore, the performance data are representative of a two-staged test that combines the use of real-time PCR and sequencing analysis of plasma EBV DNA. For example, real-time PCR assessment of plasma EBV DNA is performed as the first-line test. In total, 5.5% of tested subjects (comprising true NPC and false-positives) would have detectable levels of plasma EBV DNA. These samples would then be additionally analyzed by the sequencing test. This would represent a much more cost-effective approach because ∼95% of the population could be screened negative by the real-time PCR test.

Our work has highlighted the value of studying the fragmentation patterns of viral nucleic acids in human plasma. The clinical significance of the presence of EBV DNA in plasma of subjects without NPC is currently unknown. By showing the differences in the molecular nature of these molecules from those found in plasma of NPC patients, we offer some level of reassurance that they are less likely to represent a predisposition to NPC. Nonetheless, we are currently following up these individuals on an annual basis to assess their clinical outcome in the future. On the other hand, it would be worthwhile to explore the fragmentation patterns of plasma EBV DNA in different EBV-associated diseases or cancers, for example, infectious mononucleosis, Hodgkin lymphoma, Burkitt's lymphoma, and post-transplant lymphoproliferative disorder. Such work would be useful for establishing disease-specific molecular signatures of plasma EBV DNA and for understanding the pathophysiology of EBV in different diseases. In future studies, the fragmentation patterns of circulating DNA molecules of other viral species associated with cancers (33) could also be analyzed. For examples, circulating hepatitis B virus DNA in patients with hepatocellular carcinoma and circulating human papillomavirus DNA in patients with cervical cancer could be studied.

Early cancer detection is a challenging goal in public health and biomarker research (34). This is illustrated by a recent case control study for cancer detection through the analysis of circulating tumor-derived DNA and cancer-associated proteins (35). The reported sensitivities for stage I and II cancers based on that approach were 43% and 73%, respectively. As illustrated in our current study, a better understanding of the biological and molecular characteristics of tumor-derived circulating DNA may offer insights into further enhancing the PPV of using plasma DNA for early cancer detection.

In summary, we have developed a second-generation approach for screening of NPC. This approach is based on the differentiating quantitative and size-based characteristics of plasma EBV DNA between NPC patients and non-NPC subjects. Such an approach not only demonstrates a more superior performance in reduction of false positives but also allows a single time-point testing without the need of a follow-up blood sample. We believe that this more clinically practical protocol would greatly streamline testing and facilitate the implementation on a population scale. It is envisioned that the mortality rate from NPC would potentially be reduced as a result of mass screening in endemic regions. This study has also shed light on avenues of future developments for plasma DNA-based screening for other cancer types.

Materials and Methods

Study Design and Subject Recruitment. This study involved analysis of the plasma samples of patients with NPC and non-NPC subjects from the published prospective screening cohort (7) and an independent unscreened cohort.

The study protocol of the prospective screening study has been reported earlier (7). In brief, we organized public health education sessions in Hong Kong for subject recruitment. Our target population was asymptomatic, ethnically Chinese males aged between 40 and 62 y. This group has the highest age-specific incidence of NPC. The exclusion criteria included history of cancer or autoimmune diseases and use of systemic glucocorticoids or immunosuppressive therapy. All of the participants provided a venous blood sample of 20 mL at enrollment. Eight hundred microliters of plasma were used for EBV DNA analysis by a real-time PCR assay (8), which targeted the BamH1-W fragment of the EBV genome. The remaining plasma samples were stored at −80 °C. Any detectable level of EBV DNA by the PCR assay was regarded as a positive result. Participants who had a positive result in their baseline tests would be arranged a follow-up test ∼4 wk later using the same PCR assay. Participants who had positive results for both baseline and follow-up tests were defined as screen-positive in this study. Screen-positive subjects were referred for endoscopic evaluation and MRI of the nasopharynx for definitive diagnosis. The study was approved by the joint ethics committee of the Chinese University of Hong Kong–Hospital Authority New Territories East Cluster. All of the participants provided written informed consent for sequencing analysis of the plasma samples.

An independent cohort of patients (age range, 25–79 y) with NPC who presented symptomatically to the Department of Clinical Oncology, Prince of Wales Hospital were recruited into the current study. Twenty milliliters of peripheral venous blood were collected from each patient. All of the recruited subjects provided written informed consent for sequencing analysis of their plasma samples.

The clinical stages of all NPC patients from both the screening cohort and the independent cohort were determined based on the MRI findings according to the tumor-node-metastasis staging system of the American Joint Committee on Cancer (AJCC), as described in the seventh edition of the AJCC cancer-staging manual.

A power analysis was performed to determine the number of NPC patients and non-NPC subjects required in the validation sample set. Based on the data including the means and SDs of the EBV DNA size ratios between these two groups from the exploratory dataset, if we would reproduce the difference observed at a significant level of 0.01, a power of 0.99 would be achieved with the use of at least 200 non-NPC subjects and the remaining 24 NPC patients from the prospective study cohort in the validation cohort.

Blood Sample Collection and Plasma DNA Extraction. Peripheral blood samples were collected into EDTA-containing tubes and immediately stored at 4 °C. The blood samples were first centrifuged at 1,600 × *g* for 10 min at 4 °C, and the plasma portion was recentrifuged at 16,000 × *g* for 10 min at 4 °C to remove the residual blood cells. All of the plasma samples were stored at −80 °C until further analysis. Plasma DNA was extracted from 4 mL of plasma. DNA from plasma was extracted using the QIAamp DSP DNA Blood Mini Kit (Qiagen).

DNA Library Construction. Indexed plasma DNA libraries were constructed using the KAPA Library Preparation Kit (Kapa Biosystems) according to the manufacturer's protocol. The adaptor-ligated DNA was amplified with 13 cycles of PCR using the KAPA HiFi HotStart ReadyMix PCR Kit (KAPA Biosystems).

Target-Capture Enrichment. For enrichment of viral DNA molecules from the plasma DNA samples for subsequent sequencing analysis, target enrichment

with EBV capture probes was performed. The EBV capture probes, which covered the entire EBV genome, were ordered from Roche NimbleGen (SeqCap EZ Developer; Roche NimbleGen). DNA libraries from five samples were multiplexed in one capture reaction. Equal amounts of DNA libraries for each sample were used. We had also included probes to cover human autosomal regions for reference. The captured autosomal DNA sequences were used for normalization of the viral DNA reads. GC-neutral (i.e., with a mean GC content percentage of 40%, IQR of 34–45%) regions of human chromosomes were chosen for enrichment of autosomal DNA sequences. Since EBV DNA was a minority in the plasma DNA pool, a ~100-fold excess of EBV probes relative to the autosomal DNA probes were used in each capture reaction. After the capture reaction, the captured DNA libraries were reamplified with 14 cycles of PCR.

Sequencing of DNA Libraries. The multiplexed DNA libraries were sequenced using either the NextSeq 500 or the HiSeq 2500 Sequencing platforms (Illumina). A paired-end sequencing protocol was used, with 75 nt being sequenced from each end.

Alignment of Sequencing Data. The paired-end sequencing data were analyzed by means of the SOAP2 software (36) in the paired-end mode. The paired-end reads were aligned to the combined reference genomes including reference human genome (hg19) and EBV genome (AJ507799.2). Up to two nucleotide mismatches were allowed for the alignment of each end. Only paired-end reads with both ends uniquely aligned to the same chromosome with the correct orientation, spanning an insert size within 600 bp, were used for downstream analysis.

Statistical Analysis. Sequencing data analysis was performed by bioinformatics programs written in Perl and R languages. The Kruskal–Wallis test was used to compare the plasma EBV DNA concentrations among the NPC patients, non-NPC subjects with transiently positive EBV DNA, and non-NPC subjects with persistently positive EBV DNA in the whole screening cohort and in the exploratory and validation datasets. The Kruskal–Wallis test was used to compare the proportion of EBV DNA reads in the three groups in the exploratory and validation datasets. A value of $P < 0.05$ was considered as statistically significant.

ACKNOWLEDGMENTS. This work was supported by the Research Grants Council of the Hong Kong SAR Government under the Theme-Based Research Scheme (T12-403/15-N), the Vice Chancellor's One-Off Discretionary Fund of The Chinese University of Hong Kong (VCF2014021), and a collaborative research agreement with Grail/Cirina. Y.M.D.L. is supported by an endowed chair from the Li Ka Shing Foundation.

1. Dawson S-J, et al. (2013) Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N Engl J Med* 368:1199–1209.
2. Thierry AR, et al. (2014) Clinical validation of the detection of *KRAS* and *BRAF* mutations from circulating tumor DNA. *Nat Med* 20:430–435.
3. Forshew T, et al. (2012) Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci Transl Med* 4:136ra68.
4. Tie J, et al. (2016) Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. *Sci Transl Med* 8:346ra92.
5. Reinert T, et al. (2016) Analysis of circulating tumour DNA to monitor disease burden following colorectal cancer surgery. *Gut* 65:625–634.
6. Diaz LA, Jr, et al. (2012) The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* 486:537–540.
7. Chan KCA, et al. (2017) Analysis of plasma Epstein–Barr virus DNA to screen for nasopharyngeal cancer. *N Engl J Med* 377:513–522.
8. Lo YMD, et al. (1999) Quantitative analysis of cell-free Epstein–Barr virus DNA in plasma of patients with nasopharyngeal carcinoma. *Cancer Res* 59:1188–1191.
9. Kanakry JA, et al. (2013) Plasma Epstein–Barr virus DNA predicts outcome in advanced Hodgkin lymphoma: Correlative analysis from a large North American cooperative group trial. *Blood* 121:3547–3553.
10. Kanakry JA, et al. (2016) The clinical significance of EBV DNA in the plasma and peripheral blood mononuclear cells of patients with or without EBV diseases. *Blood* 127:2007–2017.
11. Lo YMD, et al. (1999) Quantitative and temporal correlation between circulating cell-free Epstein–Barr virus DNA and tumor recurrence in nasopharyngeal carcinoma. *Cancer Res* 59:5452–5455.
12. Leung SF, et al. (2014) Plasma Epstein–Barr viral DNA load at midpoint of radiotherapy course predicts outcome in advanced-stage nasopharyngeal carcinoma. *Ann Oncol* 25:1204–1208.
13. Hong Kong Cancer Registry (2015) Nasopharyngeal Cancer in 2015. Available at www3.ha.org.hk/cancereg/statistics.html. Accessed March 7, 2018.
14. Kanakry J, Ambinder R (2015) The biology and clinical utility of EBV monitoring in blood. *Curr Top Microbiol Immunol* 391:475–499.
15. Chan KCA, et al. (2013) Early detection of nasopharyngeal carcinoma by plasma Epstein–Barr virus DNA analysis in a surveillance program. *Cancer* 119:1838–1844.
16. Wang H-Y, et al. (2016) Cancers screening in an asymptomatic population by using multiple tumour markers. *PLoS One* 11:e0158285.
17. Chan KCA, et al. (2003) Molecular characterization of circulating EBV DNA in the plasma of nasopharyngeal carcinoma and lymphoma patients. *Cancer Res* 63: 2028–2032.
18. Chan KCA, et al. (2005) Investigation into the origin and tumoral mass correlation of plasma Epstein–Barr virus DNA in nasopharyngeal carcinoma. *Clin Chem* 51: 2192–2195.
19. Lo YMD, et al. (2010) Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci Transl Med* 2:61ra91.
20. Jiang P, et al. (2015) Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc Natl Acad Sci USA* 112:E1317–E1325.
21. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J (2016) Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell* 164:57–68.
22. Sun K, et al. (2015) Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc Natl Acad Sci USA* 112:E5503–E5512.
23. Mouliere F, Rosenfeld N (2015) Circulating tumor-derived DNA is shorter than somatic DNA in plasma. *Proc Natl Acad Sci USA* 112:3178–3179.
24. Mouliere F, et al. (2011) High fragmentation characterizes tumour-derived circulating DNA. *PLoS One* 6:e23418.
25. Underhill HR, et al. (2016) Fragment length of circulating tumor DNA. *PLoS Genet* 12: e1006162.
26. Chandrananda D, Thorne NP, Bahlo M (2015) High-resolution characterization of sequence signatures due to non-random cleavage of cell-free DNA. *BMC Med Genomics* 8:29.
27. Yu SCY, et al. (2014) Size-based molecular diagnostics using plasma DNA for noninvasive prenatal testing. *Proc Natl Acad Sci USA* 111:8583–8588.
28. Chan KCA, et al. (2004) Size distributions of maternal and fetal DNA in maternal plasma. *Clin Chem* 50:88–92.
29. Ambinder RF (2017) Plasma Epstein–Barr virus DNA for screening. *N Engl J Med* 377: 584–585.
30. Shaw JE, Levinger LF, Carter CW, Jr (1979) Nucleosomal structure of Epstein–Barr virus DNA in transformed cell lines. *J Virol* 29:657–665.
31. Cao SM, Simons MJ, Qian CN (2011) The prevalence and prevention of nasopharyngeal carcinoma in China. *Chin J Cancer* 30:114–119.
32. Sharp L, et al. (2013) Using resource modelling to inform decision making and service planning: The case of colorectal cancer screening in Ireland. *BMC Health Serv Res* 13: 105.
33. Mesri EA, Feitelson MA, Munger K (2014) Human viral oncogenesis: A cancer hallmarks analysis. *Cell Host Microbe* 15:266–282.
34. Bardelli A, Pantel K (2017) Liquid biopsies, what we do not know (yet). *Cancer Cell* 31: 172–179.
35. Cohen JD, et al. (2018) Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* 359:926–930.
36. Li R, et al. (2009) SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* 25:1966–1967.