scientific reports

Check for updates

OPEN

Preparation of robust synthetic control samples and their use in a metatranscriptomic clinical test

Ryan Toma[1,2]✉, Lan Hu[1,2], Guru Banavar[1,2] & Momchilo Vuyisich[1,2]✉

Metatranscriptomics (MT) has the potential to revolutionize the field of molecular diagnostics. Due to the complexity of MT diagnostic models, positive and negative control materials for specific disease indications can be difficult to obtain. Controls must often be sourced directly from patients. This introduces logistical burdens, assay variability, and limits high throughput clinical laboratory operations. To overcome this limitation, we developed a method for generating Synthetic Control (SC) samples, which duplicate the nucleic acid signature of complex clinical specimens and produce the desired test outcome. SCs can be easily and cost-effectively produced in large quantities (>100,000 SCs per amplification cycle), enabling high throughput diagnostic testing. Here, we report the generation of Synthetic Positive Control (SPC) samples. SPCs were validated and implemented in a clinical laboratory. The SPCs produced robust positive signals (average OC risk score of 0.996) and high levels of reproducibility (%CV of 0.29%) in a high throughput automated CLIA laboratory. SCs are a novel and useful method for the generation of high quality controls for MT-based diagnostic tests, and their adoption could herald the widespread use of MT tests in molecular diagnostics.

Metatranscriptomics (MT), the analysis of all transcripts from all organisms present in a sample, is becoming an important analytical method in the field of screening and diagnostics, offering a comprehensive platform for the interrogation of a variety of chronic and communicable diseases[1–4]. With its ability to generate large amounts of transcriptomic data, RNA sequencing has become an increasingly popular tool for identifying genetic mutations and variations[5], detecting infectious agents[1], understanding the role of human gene expression profiles in health and disease[6], and elucidating the impact of the various human microbiomes in chronic diseases[7–9]. As MT technology continues to advance, its potential for improving the accuracy and speed of disease diagnosis is rapidly increasing, making it a potentially revolutionary tool for molecular diagnostics.

MT methodologies have already been used to develop diagnostic tests for a variety of chronic diseases that would not have been possible with prior methodologies. For example, MT has been used for accurate screening of oral and throat cancer[3], type 2 diabetes[2], irritable bowel syndrome[10], and autism spectrum disorder[4]. Despite the advantages of MT in disease diagnostics there are numerous challenges that must be overcome before these tests can be translated from research into the healthcare system.

Unlike other targeted diagnostic methods such as quantitative PCR or amplicon sequencing that are looking for a few gene mutations, most MT models utilize machine learning to identify hundreds of molecular features associated with an indication[2,3,11]. While adequate positive or negative control materials can readily be generated for targeted methodologies by introducing the specific mutations or genes of interest, this is not feasible with the complex signals seen in MT disease models[12–14]. Control material is an essential component of any test as it serves as a known reference point for the performance of the test and helps to ensure accuracy and reproducibility. However, obtaining large amounts of high-quality control material for clinical assays can be challenging, particularly for unbiased tests. The shortage of control material negatively affects the clinical validation of tests and the confidence in their results, and is a barrier for the widespread implementation of MT-based diagnostic tests.

Traditional control material for sequencing based assays is typically from patients that are known to be positive or negative for an indication of interest[15,16]. However, the collection of these samples can be difficult and the sample is finite, which means that more control material from different patients must continually be obtained. This can result in significant logistical challenges. The need to continually obtain control material from different patients also raises questions about the stability, consistency, and reliability of the control source over time.

[1]Viome Research Institute, Viome Life Sciences, Inc., Seattle, WA, USA. [2]Viome Research Institute, Viome Life Sciences, Inc., New York, NY, USA. ✉email: ryan.toma@viome.com; momo@viome.com

To address these challenges, our laboratory has developed Synthetic Controls (SCs) for MT assays. SCs can be made by unbiased amplification of total RNA extracted from a clinical specimen. This process allows for the generation of a highly controlled and standardized sample that mimics the RNA profile of the original sample, thus consistently returning a known test outcome. The use of SCs eliminates the need to continually obtain specimens from patients, as they provide an almost limitless source of synthetic control material.

Herein we present the methods for generating SCs and present data from the implementation of synthetic positive controls (SPC) for an oral and throat cancer screening test in a CLIA laboratory. The results demonstrate robust SPC performance that is clinically useful. We show that SPCs are a viable and effective approach to controlling MT tests and could be adopted as standard practice.

Methods

Ethics

This study (clinicaltrials.gov registration number NCT05451303) was conducted with a protocol and consent forms approved by the Viome Institutional Review Board (IRB), an IRB accredited by the United States Health and Human Services. All experimental protocols were approved by the Viome IRB. All methods were carried out in accordance with relevant guidelines and regulations. Informed consent was obtained from all subjects and/or their legal guardian(s).

SC preparation and initial validation

Two saliva samples from patients with clinically adjudicated oral cancer (OC) were used to generate SPCs. These samples were verified to test positive using CancerDetect, an oral and throat cancer screening test that is a clinically validated and licensed Laboratory-Developed Test (Viome Life Sciences)[3]. The first participant (SPC_1) was a 59 year old hispanic/latino female that was positive for oral squamous cell carcinoma (OSCC) with a tumor node and metastases stage of T4aN2b. The second participant (SPC_2) was a 64 year old white female that was positive for OSCC in stage IV. Total nucleic acids were purified, and DNA was degraded using DNase, followed by heat inactivation. Human and microbial ribosomal RNAs were removed using subtractive hybridization. The remaining transcripts were converted to cDNA with 5' and 3' PCR primer-annealing adapters appended (Fig. 1). To preserve the integrity of the RNA, the fragmentation and cDNA size selection were not performed, however the resulting DNA profiles were not assessed. The methods for SC generation up through the end of cDNA synthesis is identical to our laboratories previously published methods (sample lysis, nucleic acid extraction, and rRNA removal)[17]. The cDNA was purified prior to PCR (AMPure XP Reagent, Beckman Coulter). The cDNA pool was amplified by PCR, where the forward primer contains a T7 promoter sequence while the reverse primer was designed to be as short as possible while maintaining a comparable melting temperature to the forward primer (Fig. 1). The resulting PCR product (F0) was used for in vitro transcription with AmpliScribe T7 High Yield Transcription Kit (Biosearch Technologies, AS3107) (Fig. 1). RNA from the transcription reaction was Dnase treated, cleaned up (AMPure XP Reagent, Beckman Coulter), and resuspended in nuclease free water.

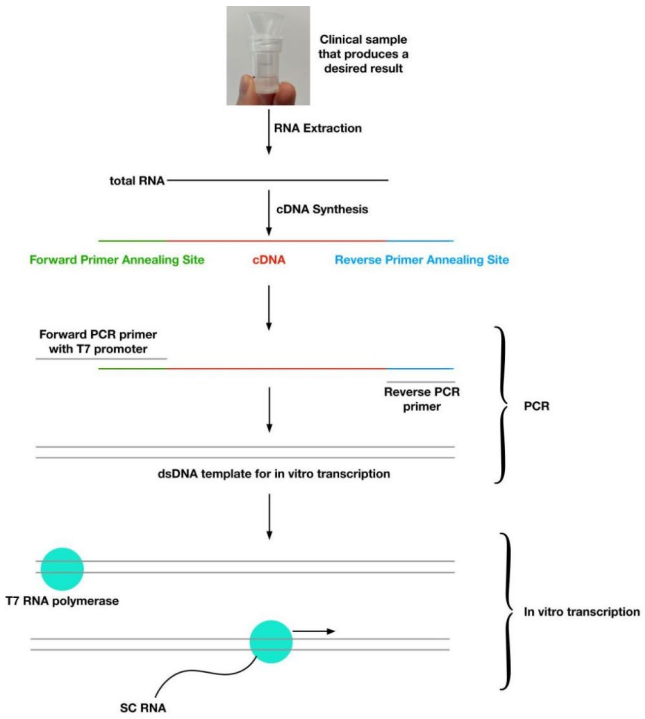


Fig. 1. Preparation method for synthetic controls.

For the test of repeated amplification, the initial PCR product (F0) was purified, quantified, diluted, and re-amplified by PCR (Fig. 2). Purified dsDNA template (3 µg) was used as input for the next PCR reaction. Each PCR step consisted of 35 PCR cycles. This process was repeated three times (F1-F3). Each resultant F0-F3 dsDNA template was also in vitro transcribed as described above. Each SPC sample (150 ng of amplified RNA) was analyzed using the CancerDetect test in triplicate. In addition, 158 samples were collected from the general population and were analyzed through the CancerDetect test as a comparison cohort (all samples had ≥ 500,000 microbial ESD).

Library preparation

SPC samples were analyzed using the CancerDetect test. The methods for this test have been previously published[17,18]. Briefly, the method includes DNase treatment, non-informative RNA depletion, cDNA synthesis, size selection, and limited cycles of PCR for adding dual unique barcodes to each sample.

SPC clinical validation

SPCs from F1 were generated for use in a clinical laboratory (Viome Life Sciences) using the methods outlined above. For the validation of SPCs, 18 replicates were processed through the CancerDetect test and their OC risk scores (also called diagnostic probability) were calculated. The OC risk score is the predicted probability returned by the oral and throat cancer classifier, pre-trained and validated as previously described[3]. SPCs used in the clinical test had to have ≥ 10,000,000 total single reads, ≥ 500,000 microbial ESD, ≥ 1,000 KO richness, and ≥ 100 species richness to pass QC and be included in downstream analysis. All 18 replicates passed these QC criteria.

Bioinformatics and disease classification

Our laboratory maintains a custom reference catalog which includes 32,599 genomes from NCBI RefSeq release 205 'complete genome' category, 4,644 representative human gut genomes of UHGG[19], ribosomal RNA (rRNA) gene sequences, and the human genome GRCh38[20]. These genomes cover archaea, bacteria, fungi, protozoa, phages, viruses, and the human host. The microbial genomes have 98,527,909 total annotated genes. Our laboratory adopts KEGG Orthology (KO)[21] to annotate the microbial gene functions using eggNOG-mapper[22].

The microbiome pipeline maps paired-end reads to this catalog using Centrifuge[23] for taxonomy classification (at any taxonomy rank). Reads mapped to the host genome and rRNA sequences are tracked for monitoring but excluded from further analysis. Reads mapped to microbial genomes are processed with an Expectation–Maximization (EM) algorithm[24] to estimate the expression level (or activity) in the sample. Respective taxonomy ranks (strains, species, genus, etc.) can be easily aggregated from the genomes. For this study, we use species activity in the downstream analyses. These genome mapped reads are extracted and mapped to only gene or open reading frame (ORF) regions for molecular function or KO annotation and quantification.

We define the number of reads mapped to the microbiome as 'microbial ESD' (Effective Sequence Depth) to represent the usable portion of reads in a sample for microbiome. ESD only refers to the reads that have an identical match to the microbial database over contiguous 240 bps, which allows clinically validated, strain level

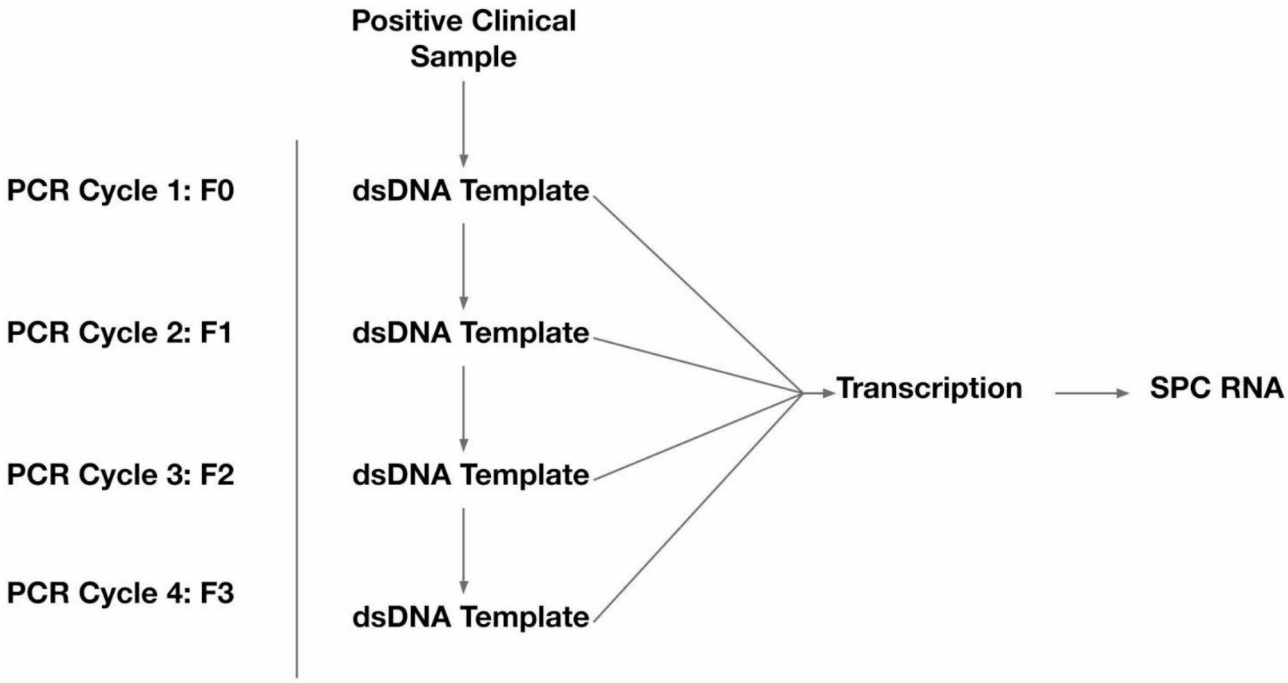


Fig. 2. Diagram of repeated dsDNA template PCR amplifications.

Amplification Cycle	dsDNA Template Yield (ng)	SPC RNA Yield per PCR Reaction (ng)
F0	453	119,590
F1	839	181,163
F2	546	311,276
F3	NA	346,968

Table 1. Average dsDNA template and RNA yields of F0-F3.

Amplification cycle	Total single reads (Av)	Total single reads (Stdev)	Microbial ESD (Av)	Microbial ESD (Stdev)	%Microbial ESD (Av)	%Microbial ESD (Stdev)
F0	9,387,282	607,089	923,981	141,391	9.84	1.17
F1	9,983,775	1,273,648	525,875	67,850	5.27	0.60
F2	7,876,085	747,307	120,500	23,281	1.53	0.20
F3	8,338,866	925,981	44,927	13,056	0.54	0.11

Table 2. Average sequencing metrics for SPC samples.

taxonomic classification. Many additional reads align to the microbial ORF (open reading frame) database. The percentage of total reads aligning to our custom reference catalog is referred to as %Microbial ESD. OC risk scores were calculated based on our laboratories previously published machine learned (ML) classifier for oral and throat cancer that requires at least 500,000 microbial ESD to return a valid result[3]. Statistical analyses were performed with Mann–Whitney U (MWU) test.

Results

SPC metrics

Two patient samples were shown to be positive for Oral or Throat Cancer via the CancerDetect test and the diagnosis was confirmed by clinical examination. The predicted OC risk score from the CancerDetect test was 0.99 for donor SPC_1 and 0.89 for donor SPC_2. Each positive sample was then converted into a SPC.

The resulting dsDNA (F0) was used as a template for in vitro transcription or purified for repeated amplification and transcription (F1-F3). The average RNA and dsDNA yields of F0-F3 are reported in Table 1. Note that the yields reported in Table 1 correspond to the output of 7 transcription reactions.

F0 generated sufficient dsDNA template for > 150 additional PCR reactions, each of which can generate > 100ug of SPC RNA. A single F0 PCR reaction is therefore capable of producing > 15 mg of F1 SPC RNA. The quantity of SPC RNA produced by the method is sufficient for > 100,000 SPC samples, demonstrating the scalability of the method.

SPC samples generated ample sequencing data (Table 2). %Microbial ESD was markedly suppressed with increased PCR cycling, indicating that repeated PCR amplifications reduces the amount of usable data. We hypothesize that the reduction in %microbial ESD with increased amplification cycles is caused by PCR artifacts of unknown composition.

SPC OC risk scores versus amplification cycles

Even though the SPC_1 and SPC_2 OC risk scores were not significantly different between F0 and F1, F2, and F3 ($p > 0.05$), there is a trend towards a reduction in the risk score with increased amplification cycles. The results indicate that SPCs should undergo a maximum of one additional round of PCR (up to F1) to minimize the degradation of the positive signal. (Fig. 3). However, this could be different for other tests. Even for this particular test, if the threshold for a positive test result is changed at some point in the future, re-amplification may or may not be appropriate.

SPC OC risk scores versus different populations

All SPCs (SPC_1 F0-F3 and SPC_2 F0-F3, combined data) showed significantly increased positive signal compared to the general population ($p < 0.001$, Fig. 4). SPCs are effective at producing robust positive signals that are significantly higher than that of the general population even with repeated PCR amplification.

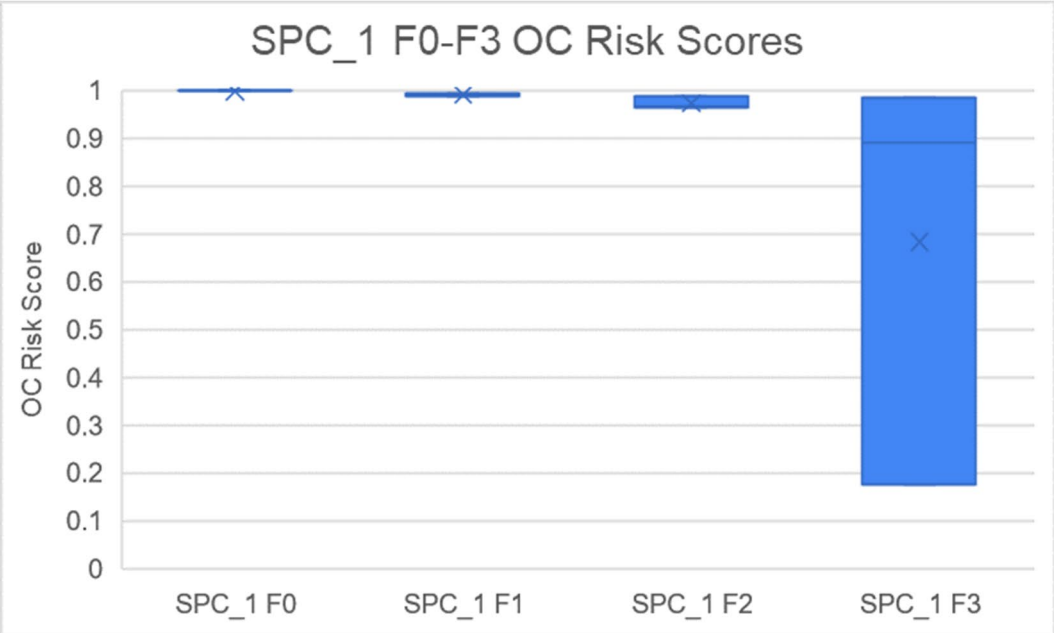
SPC implementation in a clinical test

SPCs were implemented in a clinical laboratory that performs a high throughput test for oral and throat cancer screening (CancerDetect, Viome Life Sciences). 124 SPC RNA samples generated from SPC_1 F1 were analyzed over several months, and the average SPC OC risk score was 0.996 (standard deviation: 0.003) and the percent coefficient of variation between all replicates was 0.29% (Fig. 5). These results demonstrate that SPCs can be utilized in a clinical test with high confidence in the SPC returning the desired outcome (positive).

Discussion

The utilization of MT in diagnostic tests has the potential to significantly improve healthcare. However, the lack of suitable control materials remains a significant challenge in its clinical adoption. Currently, for MT diagnostic assays with complex molecular features, controls must be obtained from patients known to be positive or negative

A.



B.

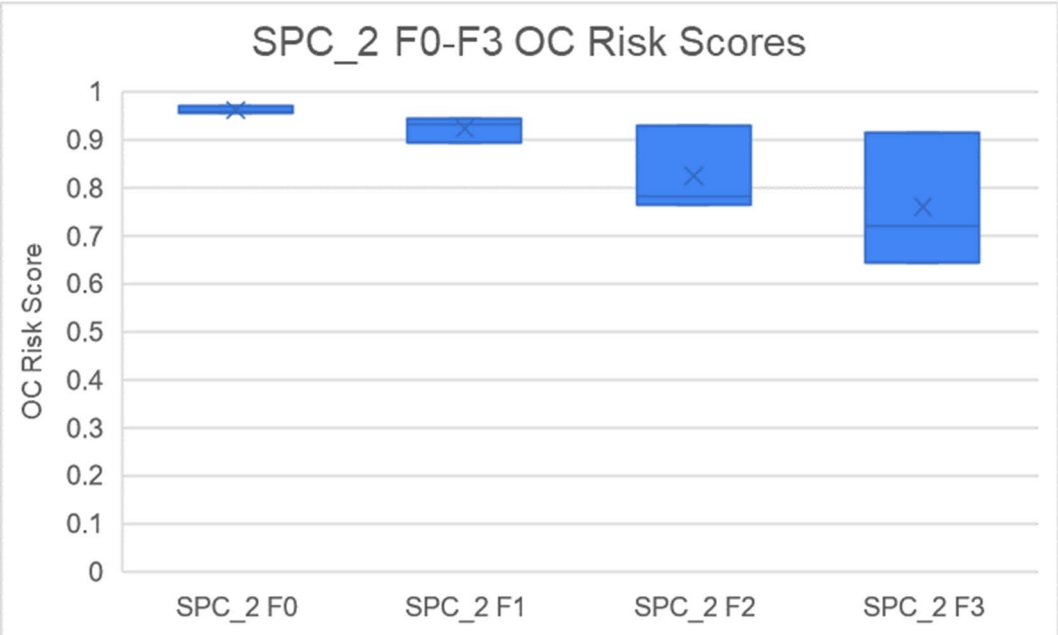


Fig. 3. SPC sample performance is affected by repeated PCR amplification. Panel A shows the performance of SPC_1, which is obtained from saliva donor 1. Panel B shows the performance of SPC_2, which is obtained from saliva donor 2. F0 is the DNA template generated by the first round of PCR, and each following round of PCR created a new version of the template (F1, F2, and F3).

for an indication. However, this method is often problematic as sample collection can be difficult and the sample amounts are limited. This results in the need to continually obtain more control material from different patients, which can be logistically difficult.

In order to address these issues, our laboratory has developed a new method for producing control materials for MT assays (Fig. 1). Synthetic Controls (SCs) can be created by amplifying the total RNA, which includes the

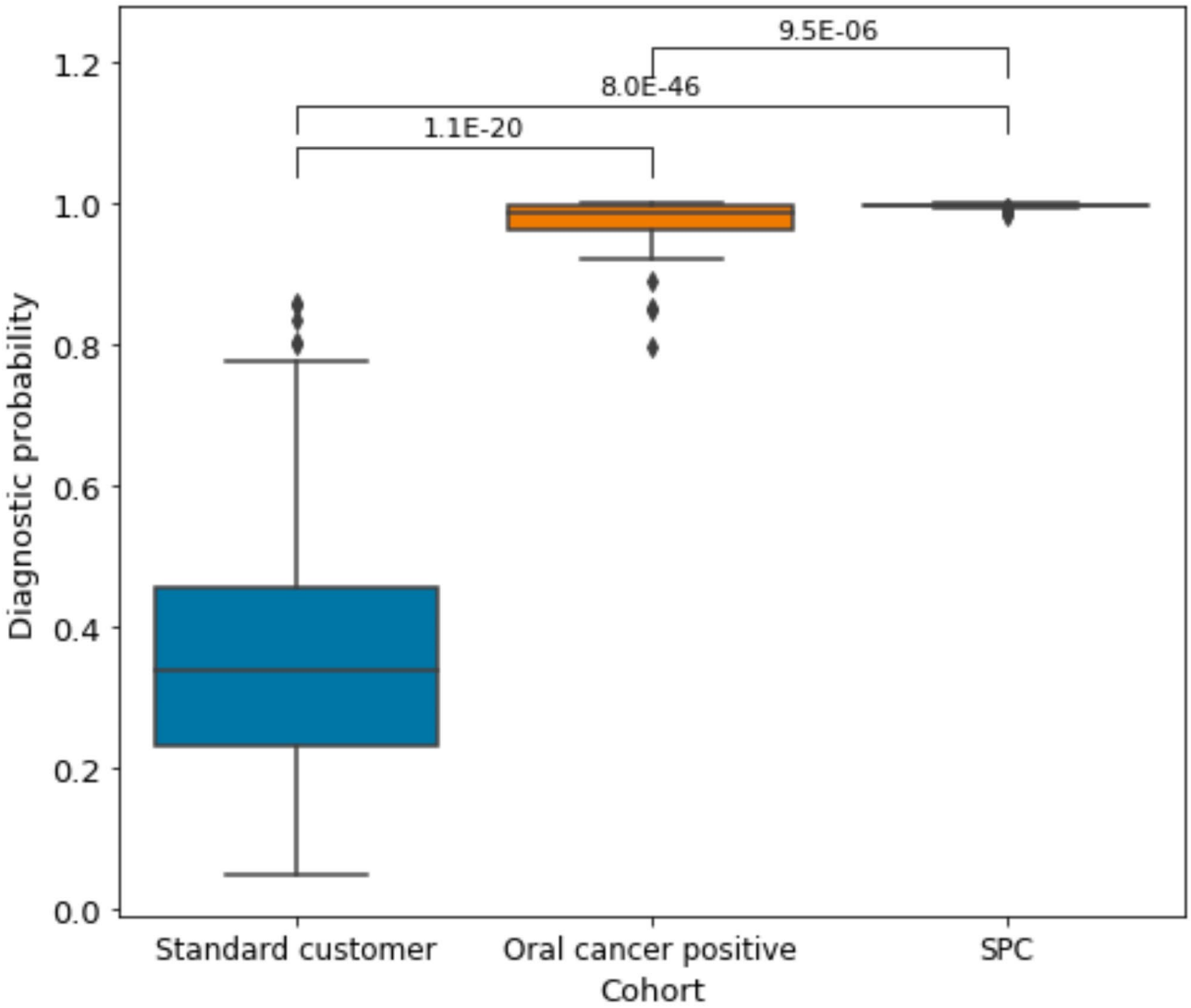


Fig. 4. SPC samples produce a very high OC risk score (diagnostic probability). Standard customer refers to the Dx probability for a set of random customers. Oral cancer positive cohort comprises saliva samples from patients diagnosed with cancer. SPC refers to the Dx probability for a set of SPC samples generated using the method reported in this study. The numbers shown at the top of the graph refer to the p values for the difference in the diagnostic probability between groups.

diagnostic signatures, from samples that are known to be positive or negative. This allows for a near limitless source of control material that is easy and cost effective to produce, and eliminates the need for continuous collection of control material from patients.

One of the key advantages of SCs is their stability and robustness. Positive SCs (SPCs) were shown to produce positive signals even after multiple rounds of DNA template amplification (Figs. 3 and 4), indicating that the positive signal is stable and that the method does not introduce significant sources of bias. This makes SPCs ideal for use as control materials in MT diagnostic tests. In addition, SPCs are also cost-effective, as they eliminate the need for frequent collection of positive control material from patients and can be produced in large quantities in a controlled laboratory setting. Future research should seek to replicate these findings with negative SCs (SNCs).

We also demonstrate a clinical application of SPCs. SPCs were utilized as positive controls in a clinical test for an oral and throat cancer screening test (CancerDetect by Viome Life Sciences). The results demonstrate that SPCs can be translated into a clinical setting while maintaining their robust performance (Fig. 5). Few research methods are ever implemented into clinical laboratories due to issues with high throughput scaling, reproducibility, and cost. SPCs overcome these challenges and have been shown to be a feasible, effective, and relevant control for an actual clinical MT test. A potential limitation of this method is that we only demonstrated its utility for one clinical test, for detection of two cancers. The methods outlined in this paper should be tested with different diseases to corroborate the method's utility across a range of relevant diseases.

The concept of SC preparation can likely be applied to additional next generation sequencing (NGS) molecular biology diagnostic techniques such as metagenomics and amplicon sequencing. Metagenomics for

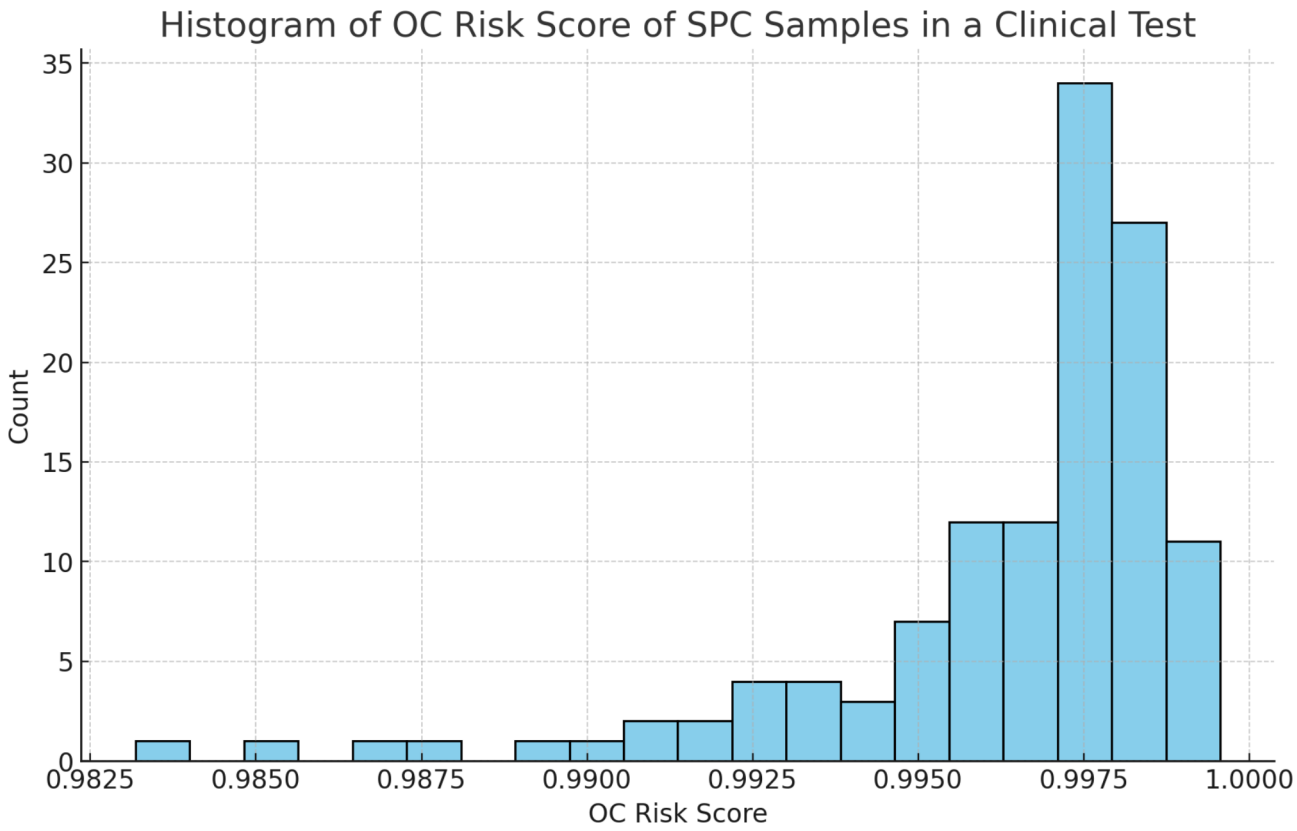


Fig. 5. SPC samples utilized in a clinical test show robust performance with a %CV of 0.29% after more than 100 analyses. Count (Y axis) refers to the number of SPC samples in each Dx probability bin (X axis).

example has already been used to generate models for a variety of conditions[12,25–28]. These NGS assays would greatly benefit from access to improved control materials. Future research should seek to apply the methods in this paper to other fields of NGS diagnostics.

SCs offer a significant advancement in the field of MT diagnostics. The stability, robustness, and cost-effectiveness of SCs make them ideal for use as control materials in MT assays. By utilizing SCs, the field of molecular diagnostics can improve the accuracy and reliability of MT and potentially other NGS tests, thereby making them a mainstay in the modern healthcare system.

Data availability

The datasets generated and/or analyzed during the current study are available in the NIH National Library of Medicine repository, under accession number: PRJNA974434.

Received: 28 April 2023; Accepted: 18 March 2025
Published online: 24 March 2025

References

1. Toma, R. et al. Pathogen detection and characterization from throat swabs using unbiased metatranscriptomic analyses. *Int. J. Infect. Dis.* **122**, 260–265 (2022).
2. Shen, N. et al. Gut microbiome activity predicts risk of type 2 diabetes and metformin control in a large human cohort. *Medrxiv* https://doi.org/10.1101/2021.08.13.21262051 (2021).
3. Banavar, G. et al. The salivary metatranscriptome as an accurate diagnostic indicator of oral cancer. *Npj Genom. Med.* **6**, 105 (2021).
4. Hicks, S. D. et al. Validation of a salivary RNA test for childhood autism spectrum disorder. *Front. Genet.* **9**, 534 (2018).
5. Chang, Y.-S. et al. Metatranscriptomic analysis of human lung metagenomes from patients with lung cancer. *Genes* **12**, 1458 (2021).
6. Marques-Coelho, D. et al. Differential transcript usage unravels gene expression alterations in Alzheimer's disease human brains. *Npj Aging Mech. Dis.* **7**, 2 (2021).
7. Vijay, A. & Valdes, A. M. Role of the gut microbiome in chronic diseases: A narrative review. *Eur. J. Clin. Nutr.* **76**, 489–501 (2022).
8. Ojala, T., Kankuri, E. & Kankainen, M. Understanding human health through metatranscriptomics. *Trends Mol. Med.* **29**, 376–389 (2023).
9. Jovel, J. et al. Metagenomics versus metatranscriptomics of the murine gut microbiome for assessing microbial metabolism during inflammation. *Front. Microbiol.* **13**, 829378 (2022).
10. Jacobs, J. P. et al. Multi-omics profiles of the intestinal microbiome in irritable bowel syndrome and its bowel habit subtypes. *Microbiome* **11**, 5 (2023).
11. Huang, Q., Zhang, X. & Hu, Z. Application of artificial intelligence modeling technology based on multi-omics in noninvasive diagnosis of inflammatory bowel disease. *J. Inflamm. Res.* **14**, 1933–1943 (2021).

12. Luthra, R., Chen, H., Roy-Chowdhuri, S. & Singh, R. R. Next-generation sequencing in clinical molecular diagnostics of cancer: Advantages and challenges. *Cancers* **7**, 2023–2036 (2015).
13. Audetat, A. et al. Analytic and clinical validation of a pan-cancer NGS liquid biopsy test for the detection of copy number amplifications, fusions and exon skipping variants. *Diagnostics* **12**, 729 (2022).
14. Cherry, P. D. et al. Abstract 247: Twist pan-cancer synthetic RNA fusion control for assay development. *Cancer Res.* **83**, 247–247 (2023).
15. Singh, A. K. et al. Detecting copy number variation in next generation sequencing data from diagnostic gene panels. *BMC Med. Genomics* **14**, 214 (2021).
16. Zanetti, A. et al. Setup and validation of a targeted next-generation sequencing approach for the diagnosis of lysosomal storage disorders. *J. Mol. Diagn.* **22**, 488–502 (2020).
17. Toma, R. et al. A clinically validated human saliva metatranscriptomic test for global systems biology studies. *Biotechniques* **74**, 31–44 (2023).
18. Hatch, A. et al. A robust metatranscriptomic technology for population-scale studies of diet, gut microbiome, and human health. *Int. J. Genomics* **2019**, 1718741 (2019).
19. Almeida, A. et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
20. Schneider, V. A. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
21. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
22. Huerta-Cepas, J. et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
23. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).
24. Henderson, N. C. & Varadhan, R. Damped anderson acceleration with restarts and monotonicity control for accelerating EM and EM-like algorithms. *J. Comput. Graph. Stat.* **28**, 834–846 (2019).
25. Kim, D. J. et al. Colorectal cancer diagnostic model utilizing metagenomic and metabolomic data of stool microbial extracellular vesicles. *Sci. Rep.* **10**, 2860 (2020).
26. Yu, J. et al. Gene mutational analysis by NGS and its clinical significance in patients with myelodysplastic syndrome and acute myeloid leukemia. *Exp. Hematol. Oncol.* **9**, 2 (2020).
27. Zaidi, A. H. et al. A blood-based circulating microbial metagenomic panel for early diagnosis and prognosis of oesophageal adenocarcinoma. *Br. J. Cancer* **127**, 2016–2024 (2022).
28. Tian, H. et al. Gut metagenome as a potential diagnostic and predictive biomarker in slow transit constipation. *Front. Med.* **8**, 777961 (2022).

Author contributions
MV and RT designed and implemented the study. MV and RT developed the SC preparation method. RT performed the molecular analyses of the SPC samples. GB and LH performed data analysis. All authors contributed to data interpretation. All authors contributed to the writing of the manuscript.

Funding
The funding was provided by Viome Life Sciences.

Declarations

Competing interests
All authors are employees of Viome Life Sciences, Inc. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Additional information
Correspondence and requests for materials should be addressed to R.T. or M.V.