

AlignDx: Enabling Automated, Cloud-Based Workflows for Streamlined Bioinformatic-Focused
Pathogen Surveillance

by

Manjot Singh Hunjan

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Science

in

Biology

Waterloo, Ontario, Canada, 2023

© Manjot Hunjan 2023

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The rising trends in infectious disease burden, alongside the recent COVID-19 pandemic, underline the need for effective public health disease mitigation strategies like pathogen surveillance. Improvements to surveillance systems can be realized by incorporating a variety of surveillance data sources such as comprehensive genomics and simpler point-of-care approaches. In this thesis, a novel bioinformatic-focused surveillance platform is presented for executing scientific workflows in cloud-based environments. The platform in question, AlignDx, addresses gaps in available surveillance systems via its modular component-based design providing security, workflow management, summary reports and data archiving. Two workflows were created and tested using this platform. First, a metagenomics next-generation sequencing workflow was developed for human pathogenic virus surveillance. Using a clinical nasopharyngeal RNA-seq test dataset, the workflow performed well in classification of severe acute respiratory syndrome coronavirus 2. Also, a lateral flow assay workflow was developed for mass automated point-of-care pathogen surveillance. Using an original test dataset of serially diluted LFA images, under controlled lighting, the workflow performed well in correctly classifying tests according to their manually curated results. Overall, the AlignDx platform is an effective system for automated surveillance applications and its constituent workflows are flexible and primed for further development.

Acknowledgements

I thank my supervisor, Dr. Andrew C. Doxey, for his passion, dedication, and guidance throughout my research.

I give thanks to my committee members Dr. Laura A. Hug, Jeremy A. Hirota and my supervisor Andrew for their insight and feedback.

Lastly, I thank the many co-op students, volunteers, and collaborators that contributed to this project: Zijjing Wu, Huagang Tan, Dayna Mikkelsen, Jennifer A. Aguiar, Nooran Abu Mazen, Briallen Lobb, Huan-Yi Shen, Ben Zwart, Kamalesh Paluru, J.D. Howell, William Le, and Linda Yang.

Dedication

I dedicate this to my loving family.

Table of Contents

Author’s Declaration ii

Abstract iii

Acknowledgements iv

Dedication v

Chapter 1 – Introduction..... 1

 1.1 Infectious Disease Surveillance..... 2

 1.1.1 Metagenomics..... 3

 1.1.2 LFAs 5

 1.2 Current Bioinformatic-Focused Surveillance Platforms 7

 1.3 Thesis Objectives and Outline - Construction of an online bioinformatic platform for automated pathogen surveillance 9

Chapter 2 – The AlignDx Platform 10

 2.1 Architecture Overview 11

 2.2 Software Stack – Implementation Details 13

 2.2.1 Web UI 13

 2.2.2 API..... 16

 2.2.3 Factory – Workflow Engine 19

 2.3 Envisioned AlignDx User Workflow 20

Chapter 3 – Genomics Workflows 22

 3.1 Base Pipeline 23

 3.2 Audience Targeted Genomic Workflows 25

 3.3 Pathogen Panels DB 26

 3.4 Wastewater – Surveillance Workflow Run 27

3.5 COVID-19 clinical swab dataset	29
3.6 Performance Evaluation	31
Chapter 4 – LFA Workflow	34
4.1 LFA Tests	35
4.2 Training Dataset	36
4.3 The VisuFlow Pipeline	37
4.4 Performance and Validation	38
4.5 Implementation of VisuFlow in The AlignDx Platform	42
Chapter 5 – Discussion and Conclusions	44
5.1 The AlignDx Platform	44
5.1.1 Challenges in Digital Surveillance	44
5.1.2 Design advantages of AlignDx	46
5.1.3 Avenues for Improvements	47
5.2 Genomics Workflows	48
5.2.1 Using diverse data sources for better surveillance outcomes	48
5.2.2 Next Steps for AlignDx Genomics Workflows	50
5.3 LFA Workflows	52
5.3.1 LFA test automation in AlignDx	52
5.3.2 Next Steps for VisuFlow	53
References	54
Appendix A – Supplementary Data	64

List of Figures

Figure 1.1 - General overview of metagenomics-based pathogen detection.....	4
Figure 1.2 - General lateral flow assay diagram.....	5
Figure 2.1 - AlignDx platform scope.	10
Figure 2.2 - Client-server architecture for AlignDx.	11
Figure 2.3 - Web UI dashboard.	14
Figure 2.4 - Example workflow report in browser view.	15
Figure 2.5 - Submission API endpoints.....	16
Figure 2.6 - Diagram of example Factory schema.	20
Figure 3.1 - Base pipeline architecture.....	23
Figure 3.2 - Targeted genomics workflows.....	25
Figure 3.3 - Snippet of the Pathogen Panels Db.....	26
Figure 3.4 - Surveillance/Lookout workflow via AlignDx client UI.	28
Figure 3.5 - Proportional abundance of SARS-CoV-2 across samples.....	30
Figure 3.6 - Example scatterplot for SARS-CoV-2 detection classification.	31
Figure 3.7 - Accuracy of SARS-CoV-2 detection on the clinical swab RNA-seq dataset using the AlignDx genomic workflow.....	33
Figure 4.1 - Overview of serially diluted LFA tests under controlled light conditions.....	35
Figure 4.2 - Schematic of image acquisition workflow.....	36
Figure 4.3 - Overview of the Visuflow pipeline.....	38
Figure 4.4 - Impact of hyperparameter variation on image binarization.	39
Figure 4.5 - ROC Curve across multiple hyperparameters for each test.	39
Figure 4.6 - Impact of hyperparameter value with and without BD Veritor.	40
Figure 4.7 - Impact of lighting temperature on image analysis algorithm performance.	41
Figure 4.8 - Visuflow workflow via AlignDx client UI.	43
Figure A.1 - Cluster map of the top 100 represented species post-human-filtering.....	65
Figure A.2 - Determination of True Positive Rate (TPR), False Positive Rate (FPR), and F1 statistic under variable block sizes with and without BD Veritor.	66

List of Tables

Table A.1 - SARS-CoV-2 100% sampling performance metrics. 67

Table A.2 - SARS-CoV-2 10% sampling performance metrics. 69

Table A.3 - SARS-CoV-2 1% sampling performance metrics. 71

Table A.4 - SARS-CoV-2 0.1% sampling performance metrics. 73

Table A.5 - SARS-CoV-2 0.01% sampling performance metrics. 75

Table A.6 - SARS-CoV-2 10 million read sampling performance metrics. 76

Table A.7 - SARS-CoV-2 1 million read sampling performance metrics. 78

Table A.8 - SARS-CoV-2 100 thousand read sampling performance metrics. 80

Table A.9 - SARS-CoV-2 10 thousand read sampling performance metrics. 82

Table A.10 - SARS-CoV-2 1 thousand read sampling performance metrics. 83

Table A.11 - ICON DS Dry performance metrics for block size. 84

Table A.12 - ICON DS Wet performance metrics for block size. 86

Table A.13 - QUIDEL Dry performance metrics for block size. 88

Table A.14 - QUIDEL Wet performance metrics for block size. 90

Chapter 1 – Introduction

Waves of infectious diseases carried devastating consequences throughout human history. Despite significant efforts to combat infectious diseases through vaccination, sanitation and other public health strategies, the threat of emerging infectious diseases (EIDs) and re-emerging infectious diseases (REIDs) remains [1,2]. One useful metric for understanding the cost of illness is the global burden of disease. Typically, it is calculated using measurements of population health according to the effects of disability/disease and death, represented as disability-adjusted life years (DALYs) [3]. According to the 2019 global burden of disease (GBD) report, a comprehensive assessment of health and injuries for 204 countries and territories, six infectious diseases (lower respiratory infections, diarrheal, malaria, meningitis, whooping cough and sexually transmitted infections) were among the top causes of DALYs [4]. Global disease surveys such as the GBD and the World Health Statistics report point to communicable diseases amongst the principle purveyors of disease burden in low to middle-income countries [4,5].

Technological change has caused unprecedented disturbances, leading to dramatic differences in human interconnectivity, farming practices, climate conditions, etc. [1]. This, alongside increased population sizes within human-dominated ecosystems, has magnified the human-animal interface increasing the risk of zoonotic EIDs [6]. EIDs most commonly originate via zoonosis (estimated to be 75% of EIDs) and are trending upwards in incidence [7,8]. The impact of the novel Coronavirus Disease 2019 (COVID-19) pandemic on global healthcare systems supports these trends. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the prevailing causative agent of COVID-19, is an example of an EID, spread primarily as a respiratory infection [9].

Amongst the most common infectious diseases are upper respiratory tract infections (URI), with global incidence having increased from 1990 to 2019, accounting for 43% of the GBD [10]. Similarly, the patterns in the burden of lower respiratory tract infections (LRI) have increased globally [11]. Efforts have been made to estimate the global burden of COVID-19 and while these are varied across countries, overall mortality contributes significantly to the global burden [12]. Recent studies have shown how the pandemic disrupted urgent care, leading to larger consequences in the healthcare system [13]. Given these trends, there is a clear need for concerted efforts towards disease control. Numerous disease containment strategies exist, and they can be placed within the scopes of therapeutic countermeasures and public health interventions [14]. To mitigate the burden of infectious diseases, as

well as the risk of potential pandemics, early public intervention strategies such as surveillance are essential.

1.1 Infectious Disease Surveillance

Surveillance concerns the collection, analysis and interpretation of health data, acting as an intervention or disease control strategy [15]. Infectious disease surveillance methods vary based on the effective goal, such as those of population, aggregation, syndromic and laboratory-confirmed surveillance, amongst others [16]. Modern surveillance systems make use of a variety of lab techniques, and incorporate multiple data sources for disease monitoring via the internet and computer systems for digital epidemiology [17].

Traditionally, techniques such as microscopy, culture-based methods, and serology were commonly used for pathogen diagnostics. However, in the past few decades, there has been a transition to molecular technologies, including nucleic acid amplification tests (NAATs) [18], lateral flow assays (LFAs), and increasingly, although less common, genomics-based methods. Polymerase chain reaction (PCR) and other NAATs are a gold standard in pathogen diagnostics [19]. PCR for example, can be very effective at detecting low quantities of DNA/RNA [19], and in pathogen detection, can exhibit a high sensitivity and specificity [20]. However, NAATs are limited in simultaneous identification of multiple species [20] and they rely on primers that recognize known diagnostic sequences within a target genome [19]. In a recent study on reverse transcription-polymerase chain reaction (RT-PCR) screening of SARS-CoV-2 variants highlighted limitations in following viral evolution [21]. These methods may fail to detect newly emerging pathogen variants within an acceptable time frame for a public health response. Furthermore, as these approaches are targeted, these methods are incapable of detecting novel pathogens or EIDs.

Below, two technologies for pathogen detection in the context of infectious disease surveillance are explored: metagenomics via high throughput sequencing technologies and lateral flow immunoassays (LFIAs). While these are not the only technologies to exist for pathogen detection, these are specifically explored as they represent two extremes of a spectrum in terms of their resolution, speed, and use cases. Multiple pathogens, whether EID or REID, can be detected at a high resolution with metagenomics. Although it is currently a relatively slow and expensive method, it can produce an enormous amount of data. On the other end are point-of-care (POC) technologies like LFIAs, which are widely available, cost-effective, and rapid [22]. These provide single pathogen detection within minutes at the cost of

accuracy and little diagnostic information. In infectious disease surveillance, high throughput sequencing technologies represent a comprehensive future approach to pathogen detection, whereas LFIA technology is a simple but modern alternative.

1.1.1 Metagenomics

Next-generation sequencing (NGS) based molecular strategies are quite commonly developed and put into use by individual laboratories for diagnostic purposes [19]. Metagenomics then combines these single-diagnostic methods to analyze all the genetic material from a patient sample for a variety of disease markers [23]. The diagnostic value of metagenomics is in detecting all potential pathogens within a sample without bias or reliance on culture [24]. It thus offers several advantages over traditional pathogen detection, due to its capabilities in detecting novel pathogen variants and species. Like other NGS based sequencing strategies, metagenomics involves a complex series of steps. Once a clinical specimen is sampled, DNA/RNA is extracted and purified to undergo massive parallel sequencing, and then subsequent bioinformatic analyses [25] (Figure 1.1). In terms of sequencing platforms, current metagenomic strategies typically favor short-read NGS sequencing platforms (e.g. Illumina), although there is growing use of long-read platforms (e.g. Oxford Nanopore Technology (ONT)) [26]. Pearman and colleagues compared long-read vs short-read eukaryotic metagenomics to confirm the error-prone issues of long-read technology, but also concluded that taxonomic group of interest impacts the methodology [27]. This suggests that with a targeted approach, a wider set of NGS platforms may be viable. However, the reality is that for targeted approaches, there are more cost-effective alternatives with shorter turn-around times. In contrast to traditional differential diagnostics that are hypothesis-driven, untargeted metagenomics is hypothesis-free [23]. The untargeted approach of metagenomics is therefore powerful in surveillance, or generally any early diagnostics, when no specific causative agent is suspected.

Gardy and Loman [17] suggested a surveillance model which combines syndromic and localized surveillance, where sequencing data is integral to pathogen identification. While this may be feasible at the scale of an individual laboratory, wide-scale adoption for surveillance poses massive challenges. This includes but is not limited to access, sample collection, nucleic acid extraction, library preparation, sequencing, analysis and reporting [23]. In addition, bioinformatic pipelines require complex knowledge of programming, computer infrastructure and expertise in a variety of algorithms for specific analyses [23]. Kraken is an example of a rapid metagenomic classification tool well-suited for

a clinical metagenomics pipeline [28]. However, domain-specific knowledge is required to translate its predictions into biologically/clinically meaningful results. Thus, new methods are needed to simplify bioinformatic workflows and associated prediction reports, such as through automated cloud-based systems that require minimal technical input by the user.

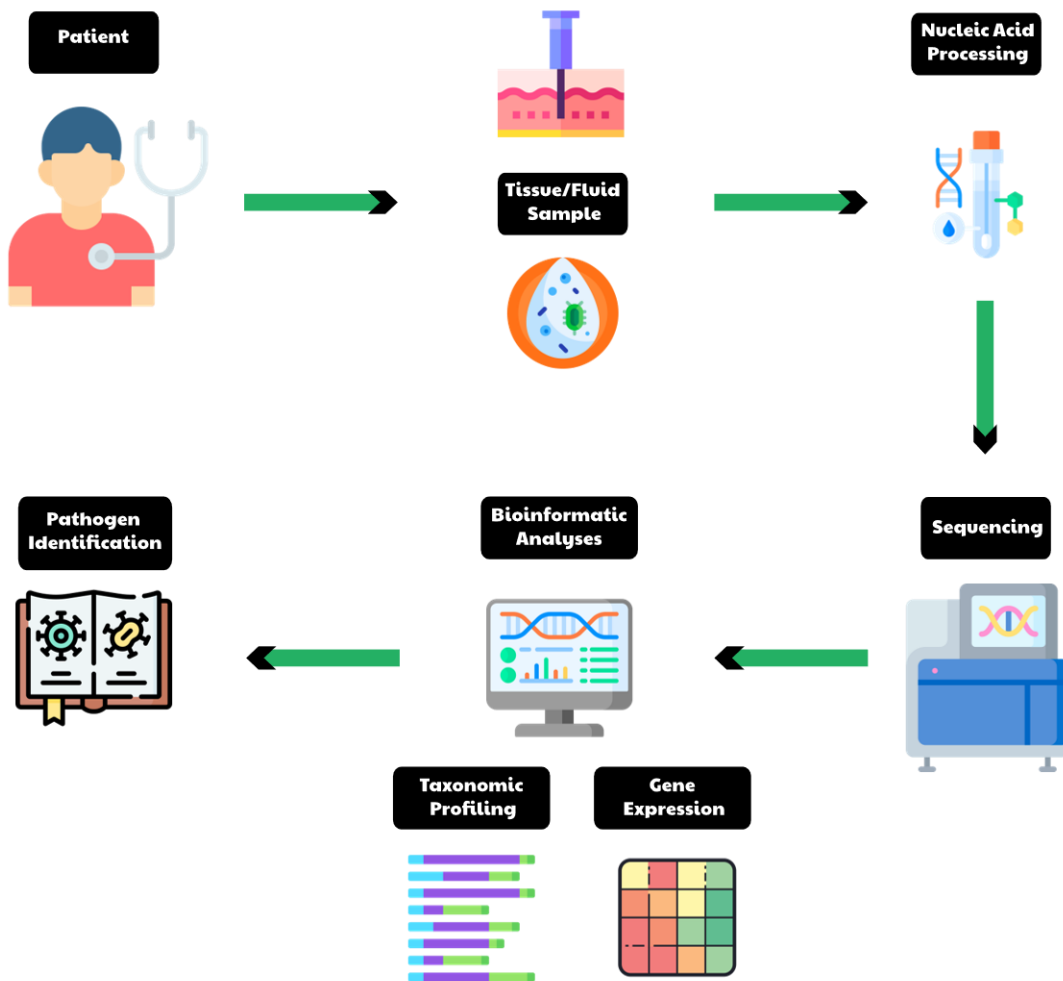


Figure 1.1 - General overview of metagenomics-based pathogen detection. A patient, with or without symptoms, is targeted for this molecular workflow. Samples are taken from areas of interest. Nucleic acids are extracted, isolated, and processed for genome sequencing. Finally, various bioinformatic workflows are run to aid in the identification of a suspect pathogen. Adapted icons from Flaticon and Icons8, and vectors from nf-core pipeline components [29].

1.1.2 LFAs

LFAs consist of a paper-based assay that enables rapid quantification of a biomolecule, classified by their recognition element as either antibody-based or nucleic acid-based [30]. Immunodiagnosics POC tests like LFAs are a widely adopted approach, typically used within the confines of routine clinical microbiology [31]. Typically, the configuration of an LFA involves a membrane strip embedded into several pads and a backing plate, allowing a sample liquid to travel via capillary action to a zone of conjugates and then detection to produce a visual response [30] (Figure 1.2). The visual response can then be compared to a control line [30].

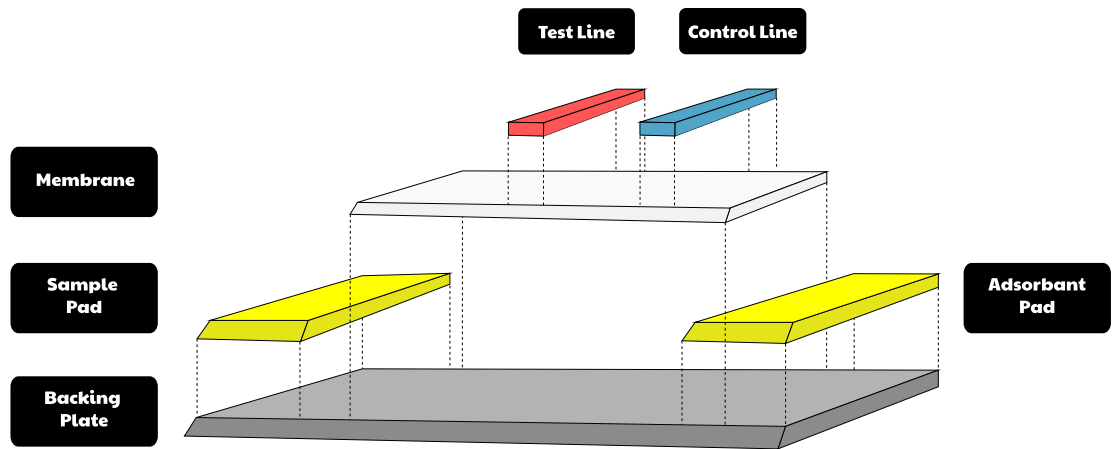


Figure 1.2 - General lateral flow assay diagram. Samples flow from the left sample pad to the right absorbent pad, passing the test and control lines through membrane material.

In the past few years, there has been a wide-scale adoption of LFIA tests for clinical and at-home screening of SARS-CoV-2 infections. Several studies have demonstrated highly sensitive and specific detection of SARS-CoV-2 antigens using these tests [32]. Their categorization as a POC test makes them ideal for mass-testing, particularly in having a short-turnaround time, cost-effectiveness and reasonable accuracy [33]. However, there are several limitations when using LFAs. One significant challenge is that test performance is largely operator dependent [33]. Relying on the subjective interpretation of the human eye can lead to varying results. Quality data on the recording accuracy of the human eye is scarce due to a lack of empirical data, but some details regarding diversity in color perception do underline this subjectivity [34,35]. Furthermore, correct usage of the LFA requires a correct understanding of their non-standardized manufacturer’s manual. Another significant challenge

is that few are FDA-approved, with great variance in which tests show acceptable sensitivity and specificity [32,33]. Perhaps the most-overlooked limitation is the lack of automation and in general, documentation of results. These limitations could in part be addressed using a central cloud-based surveillance framework to process, store and share results of LFA tests.

1.2 Current Bioinformatic-Focused Surveillance Platforms

While POC and NGS approaches are promising data sources for digital epidemiology, end-to-end surveillance pipelines incorporating these approaches are uncommon. This is in part because these approaches often require skilled technicians or bioinformaticians to have a full understanding of these domain specific workflows. For routine adoption, a system that abstracts complexities, and eases the analysis and interpretation components of the surveillance pipeline is essential. Efforts at generating such an end-to-end surveillance pipeline are rapidly developing, especially in digital epidemiology.

Platforms are in development for POC LFA approaches. One study proposed REASSURED (Real-time connectivity, Ease of specimen collection, Affordable, Sensitive, Specific, User-friendly, Rapid, Equipment-free, and Deliverable) diagnostics, in which a machine learning (ML) strategy is implemented alongside the use of smartphone technology to produce a robust, manageable system [36]. Such a model could be enhanced via a user-interface, connected to an archiving system. LFA App is an example of an open-source smartphone-based system for quantitative analysis of LFA that attempts this, but is focused on singular usage, thus it is not suited to wide-scale surveillance deployment [37]. Sequencing approaches have seen much more usage and implementation, although with different barriers. Galaxy for example, an open-source collective integrating multiple, well-established bioinformatic tools in a cloud environment, has been repurposed for clinical analyses [38]. It is primarily built as a workbench, thus is more suited to educational and research-focused workflows [39]. In contrast, Nextstrain is a real-time pathogen tracking platform, built for monitoring viral outbreaks [40]. Although effective, it is focused largely on phylodynamics, and does not implement other types of upstream bioinformatic analyses required for initial pathogen detection [40].

Bioinformatic-focused surveillance requires a personalized approach prioritizing a simple user interface (UI), focused pipelines, security, and scalability [41]. User-centered design (UCD) is also often ignored within the scope of bioinformatic tools and workflows, often due to its complexity [42]. Typically, there is a steep learning curve for using these tools, and interfaces are bustling with tweakable parameters that may be unnecessary for usage. A UCD approach could be valuable in generating a domain-specific UI that is minimally viable and allows for reproducibility [43]. Bioinformatic workflows are also highly dynamic, as tools are ever-changing, so an abstraction layer is typically required. Workflow managers are a popular solution as they provide data provenance, portability of pipelines, scalability and reentrancy of failed executions [44]. Containerization is the driving technology behind these workflow managers, which allows the packaging of workflow

processes into modular, portable, units that can be orchestrated at any desired scale [44]. This system is certainly not complete without an archive/database, that can store, and manipulate analysis results for collection and further interpretation. Numerous technologies exist, but relational database models are well suited to strongly structured data, such as those generated by bioinformatic workflows [45]. Finally, the individual components of these systems in unison must be scalable, and widely accessible. In terms of scalability, cloud computing has revolutionized access to computational resources, and is in regular use in every major tech adjacent industry. The internet is also widely accessible; thus, it acts as an excellent interface for any potential bioinformatics platform.

1.3 Thesis Objectives and Outline - Construction of an online bioinformatic platform for automated pathogen surveillance

As described above, there is a need for a computational infectious disease surveillance platform with features that address the limitations of existing systems including: user-centered design, flexibility in technology implementation, data inputs and workflows, method automation, online data archiving, security and scalability. The following chapters describe the development and testing of a new system called AlignDx, a cloud-based, user-friendly web platform for surveillance, via modern imaging and sequencing-based bioinformatic approaches. In Chapter 2, the methods for the construction of the AlignDx platform are explored, with a focus on the system architecture, and its implementation. Chapter 3 describes the AlignDx genomic workflow in detail, alongside test cases and applications using original datasets. Chapter 4 details the AlignDx LFA image analysis workflow similarly to the genomic workflows, as well as accuracy testing via a dataset of LFA images. Finally, Chapter 5 explores the AlignDx platform and its workflows.

Chapter 2 – The AlignDx Platform

Modern surveillance systems adopting digital epidemiology could greatly benefit from a bioinformatics-focused approach to disease control. Current systems face limitations such as tightly coupled data sources, narrow target audiences, and unintuitive, complex user interfaces. Herein, the construction of “AlignDx”, a cloud-based web platform for running bioinformatics-focused surveillance workflows is outlined. This system is designed to address the current gaps in surveillance systems, through a simple UI, returning consistent, workflow-specific reports (Figure 2.1).

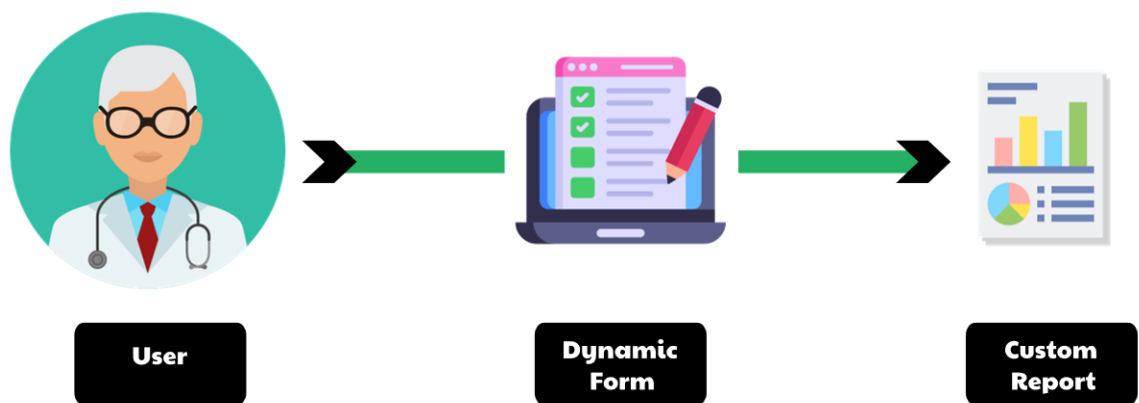


Figure 2.1 - AlignDx platform scope. Users access workflow specific forms through a dynamic UI which returns a summary report that highlights relevant results. Adapted icons from Flaticon and vectors from nf-core pipeline components [29].

2.1 Architecture Overview

Development of the AlignDx platform began with the construction of a high-level client-server architecture, shown in Figure 2.2. Through this architecture, users are expected to interact with the system using any web browser, via a web-UI client, that allows them to run curated bioinformatic workflows on a remote server. Workflow outputs are then used to generate workflow-specific reports, summarizing key findings with pertinent visualizations and descriptions. Underneath lies a service-oriented software stack, where major components are split into container-based operators running on a cloud-server.

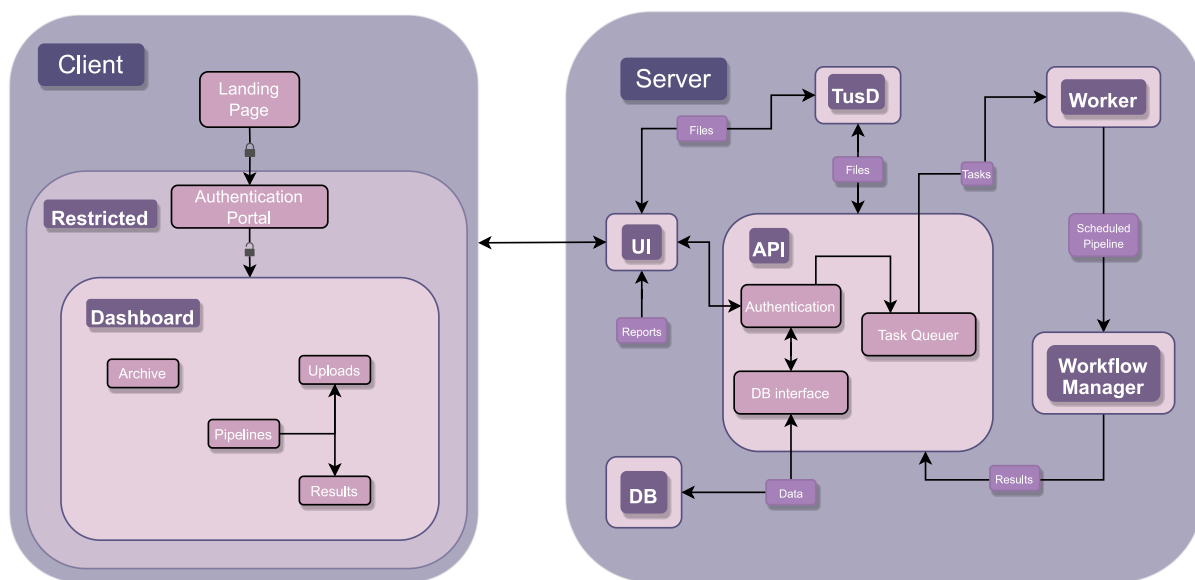


Figure 2.2 - Client-server architecture for AlignDx. Representative diagram for the overall architecture and communication pathways of system services. Client architecture is generated by the UI service on a remote server, and then sent to the client to be displayed in browser. DB: Database, UI: User Interface, API: Application Programming Interface.

Components can be grouped according to their core usage, being the front-facing web UI, and the rear, application programming interface (API) for orchestrating surveillance tasks. The UI is a client web application with a responsive design capable of running on any modern browser. Users can select a workflow and fill out its form with the required data, which is sent to the API on a remote server. The UI further enables monitoring of workflow execution, and finally, exploration of workflow results. The construction, execution and monitoring of workflows are managed internally by the API running on the remote server. This component acts as a middleman, coordinating user requests from the frontend UI to the corresponding service.

From a surveillance perspective, core features provided by the API include authentication, task scheduling, archiving, workflow management and report generation. The authentication service provides data privacy essential when dealing with sensitive information such as health data. Task scheduling ensures that massive datasets, alongside users, do not overwhelm the system. This is critical to handling varying levels of traffic. Archiving tracks submission metadata, providing a historical record of previous submissions that is fully controllable by the user. Workflow management is a function of previous components and enables the execution of any data pipeline using containerization. Report generation summarizes the findings of a specific workflow.

2.2 Software Stack – Implementation Details

The system is split into Docker [46] container-based services running on a cloud-server, described below. Each service has its own set of software dependencies and build requirements defined using the Dockerfile text file format [46]. From these text files, an executable package representing a snapshot or “image” of the service can be constructed, and then run as a process/container. All services can then be managed using an orchestration software, such as the Docker Compose tool [47], which controls service lifecycles and communication through a Compose configuration file. More advanced orchestration solutions that offer computer cluster-level deployment can also be used, due to the ubiquity of Docker containerization technology.

2.2.1 Web UI

The UI was built using the React.js [48] library and the Next.js [49] framework following UCD. The source code follows the functional programming paradigm, implemented using the TypeScript programming language for maintainability. This UI is server-side rendered and updated when necessary for rapid-response times. Rendering refers to the process by which the source code is converted to an interactive web page. By performing this server-side, significant computational load is taken away from the user. The code loosely follows atomic design methodology, where simple components serve as building blocks for larger and more complex components. The resulting components can be loosely categorized as either stateless or stateful units. Stateless components require no inputs, whether from users, or system processes such as API requests, making up the surface layer of the UI. Stateful components do the bulk of functional processing, relaying user inputs, as well as system responses back to the web client. This allows for rapid adoption of novel web technologies that may supersede current implementations. Stateful and stateless components are then combined to generate complex features, such as the dynamic form components. These then come together in page-spanning layouts that map them as required, making up an application dashboard. Each component is designed to be easily customizable and can be carried over to any React-based framework. Additionally, they are responsive, meaning their dimensions are dynamically adjusted to available screen space.

As a data-centric platform, an application dashboard ensures data-driven decision-making, where tooling is at the forefront of the users’ capabilities. Currently, the dashboard is split into two core pages: the “Analyze”, and “Archive” pages (Figure 2.3).

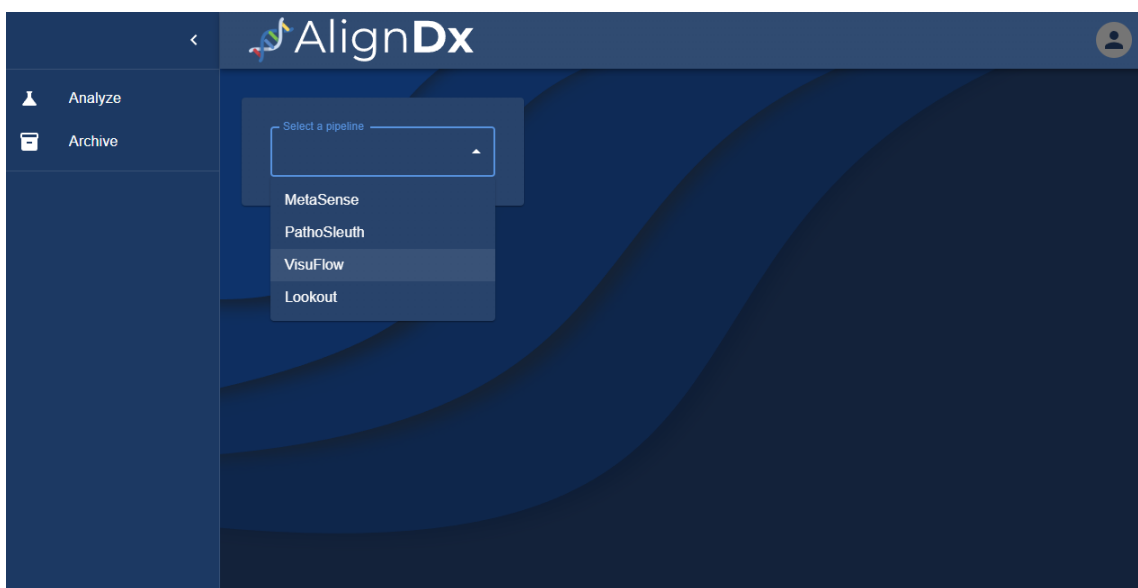


Figure 2.3 - Web UI dashboard. Central application dashboard with buttons to access “Analyze” and “Archive” pages. Shown is the “Analyze” page, with the pipeline/workflow selection menu.

The analyze page consists of several core components, key among them being the pipeline form and monitor component. The pipeline form component serves to dynamically assemble UI workflow forms as requested by the user upon selection. Forms are constructed based on keyword input declarations in the JavaScript Object Notation (JSON) [50] schema provided by the API. This includes meta-field descriptors, which can provide the user with necessary information on input requirements. Each input field also has a dynamic validation function, where incorrect or missing required inputs prevent submission, and consequent errors are visually highlighted to the user. In this manner, users receive immediate feedback through metadata, as well as form interaction. The monitor component enables a user to visually identify submission progress through status indicators, and progress bars. Progress data is fetched in real-time between the client web UI and the server API using the WebSocket communication protocol. Status updates follow a subset of workflow event triggers such as submission setup, analysis, completion, and error. Progress bars are however reserved for file data uploading and provide resumable functionality through a pause/play button. This component is given priority over other UI, to ensure that in the event of a network failure, users are immediately prompted to resume interrupted submission attempts. The archive page consists of a table component that enables exploration of previous submissions to the user. The metadata includes the submission name, selected pipeline, timestamps, and status of the submission. Additionally, each submission’s output data and

corresponding report can be accessed here. Data, including reports are downloadable in a compressed format from any device; similarly reports can be viewed in browser (Figure 2.4).

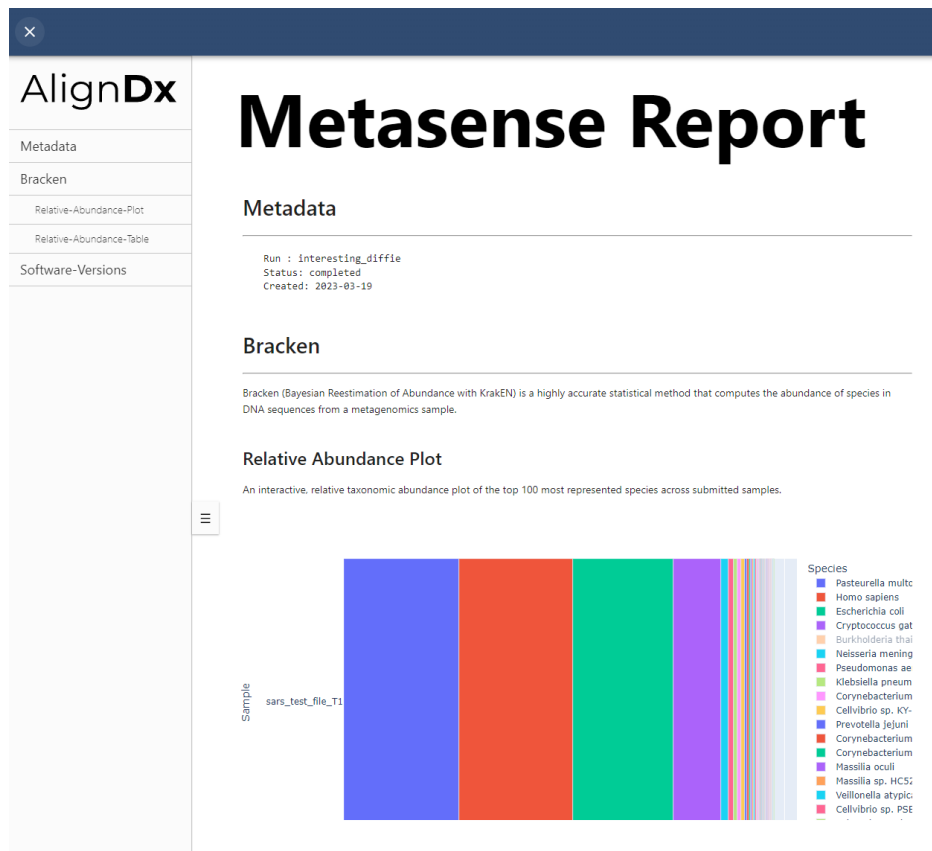
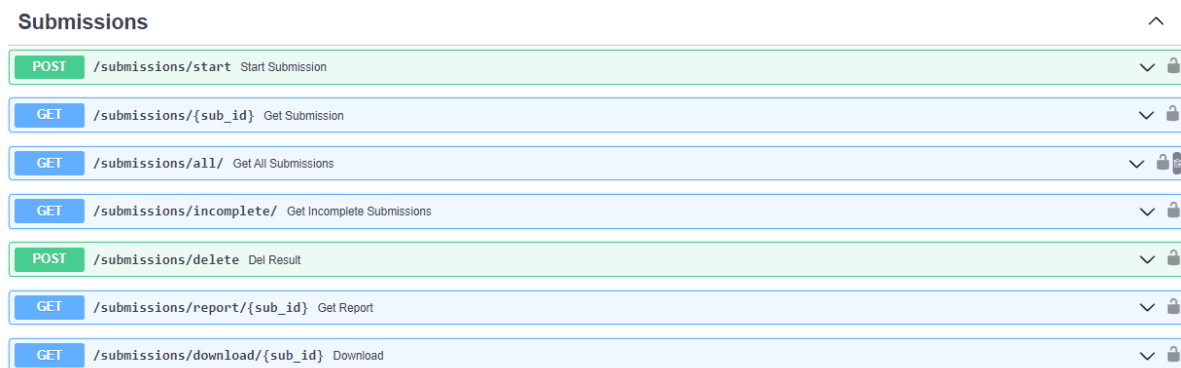


Figure 2.4 - Example workflow report in browser view. The report was generated using the MetaSense taxonomic profiling workflow, with an Ion Torrent SARS-CoV-2 positive FASTQ sample. The top 100 most represented species are visualized in the shown relative abundance plot, generated from the workflow output.

2.2.2 API

At the core of the system lies an API, built using the asynchronous FastAPI [51] web framework, to direct incoming and outgoing traffic to the appropriate service. An API can be thought of as a set of rules defining how two or more computers may communicate, such as over the web. Asynchronous here refers to the execution nature of the API, where incoming requests can overlap, and are scheduled as necessary. This translates to better performance and scalability for the client, since content can be handled dynamically. Like the web UI, the API design also follows atomic design methodology, where smaller units are used to build up more complex services. At the lowest level, there are asynchronous functional units, that provide a singular purpose. These functions are then combined with data models into a programmatic object to generate a service. Data models can be highly complex, or simple representations of incoming and outgoing data. Using the Pydantic python library [52], these are created, providing model validation and an intentional structure to service execution, which feed into system monitoring and execution control. At the highest level, these services can then work together to create complex functionality, accessible through API endpoints (Figure 2.5).



The image shows a screenshot of the Swagger UI for the 'Submissions' API. The title 'Submissions' is at the top left with an expand/collapse icon on the right. Below the title is a list of seven API endpoints, each in a colored box with a dropdown arrow and a lock icon on the right. The endpoints are: 1. POST /submissions/start Start Submission (green box, lock icon); 2. GET /submissions/{sub_id} Get Submission (blue box, lock icon); 3. GET /submissions/all/ Get All Submissions (blue box, lock icon); 4. GET /submissions/incomplete/ Get Incomplete Submissions (blue box, lock icon); 5. POST /submissions/delete Del Result (green box, lock icon); 6. GET /submissions/report/{sub_id} Get Report (blue box, lock icon); 7. GET /submissions/download/{sub_id} Download (blue box, lock icon).

Method	Endpoint	Description	Authorization
POST	/submissions/start	Start Submission	Scoped
GET	/submissions/{sub_id}	Get Submission	Scoped
GET	/submissions/all/	Get All Submissions	Scoped
GET	/submissions/incomplete/	Get Incomplete Submissions	Scoped
POST	/submissions/delete	Del Result	Scoped
GET	/submissions/report/{sub_id}	Get Report	Scoped
GET	/submissions/download/{sub_id}	Download	Scoped

Figure 2.5 - Submission API endpoints. Submission endpoints visualized in swagger UI interactive documentation, following the OpenAPI specification. At the left, request methods (GET, POST) are highlighted. At the right, locks indicate authorization scoped endpoints.

Web API endpoints are communication pathways for a particular service, and each endpoint can have a set of rules about inputs, outputs, and communication protocol. The implementation of these endpoints follows representational state transfer (REST) design principles. In short, each endpoint has a uniform interface, is stateless, cacheable, decoupled from the client and consists of communication layers [53]. This allows for flexible implementation of services, and ensures that client-server communication is predictable, making the overall system robust. The API utilizes common HTTP request methods for communication including the GET, POST and DELETE methods. These define the type of action carried out by the endpoint; a GET request retrieves data, whereas a POST request submits data, and DELETE removes data [54]. Each endpoint is organized by a common use case, as seen in figure 2.5, including metadata, user, submission, socket and finally webhook requests. Metadata endpoints inform the client on dynamic information created by the server, including analytics or in the case of this system, available workflows. User endpoints invoke authentication services, which enable or disable access to other endpoints, depending on access granted to the requestee. Simple examples include registering a user or signing in to access previous work. Submission endpoints are used to trigger workflows and retrieve results, including reports and other output data. Unlike other endpoints, socket and webhook routes differ in communication protocols, thus they serve different purposes. In the case of socket endpoints, these provide real-time updates on the state of workflows using the WebSocket protocol. Webhooks provide similar functionality but use the HTTP protocol to allow communication between third-party services and the system API.

The flexibility of the API design gives way to the integration of any feature enhancing third party-tool. One key element missing from this system was fault tolerance for network failures, especially when working with large files for bioinformatic workflows. As a result, data uploads are managed through the tus resumable protocol, via a tusD server [55]. Briefly, “resumability” provides a mechanism by which partial uploads may be continued by the user when network stability has returned. Webhooks provide the communication layer between upload events and the API, meaning the system can track when data is ready or in stasis. Furthermore, once uploads for a particular submission have been completed, the system works independently of users’ network availability for analysis of submitted data.

Every service implemented by this API makes use of 2 types of storage: disk and database (DB). These two storage types are used in sync to accommodate different purposes, where data storage can be organized by size and longevity. Files ranging between fractions to hundreds of gigabytes (GB) are common data inputs/outputs for bioinformatic workflows. Long-term storage with this data can be tricky in terms of cost, performance, and security, thus these are considered low priority in terms of storage for the system. Most of the other data input by users, namely form data or user information can be structured, easily predicted and relatively small. The system combines disk and DB storage to manage all these data types using the relational PostgreSQL DB management system [56] and the in-memory Redis DB [57]. The PostgreSQL DB provides long-term stable storage, while Redis provides rapid-access temporary storage. For file data, relevant file metadata and workflow tracking information will be temporarily stored in the Redis database and upon completion, a subset of data will be kept long-term in PostgreSQL. Form data, user data and any of their kin can be directly stored in PostgreSQL. As a relational DB, data groups can be categorized as “tables”, with rules mapping relations between groups. In this manner, user information can be associated with their submissions, and in turn any cascading data. The DB is accessed using a DB interface that is framework agnostic, meaning one could easily swap out DB paradigms if performance or technological creep necessitates it. DB querying is used extensively in key services, including authentication and workflow management, amongst others. Querying, much like network requests with the API, is done asynchronously between the API and the PostgreSQL database. Similarly, this increases the efficiency of the system, as resources can be allocated as necessary, without blocking the execution of other tasks. The querying methods are shared between tables, as defined by the DB interface. Each table can, however, have different structures defining the data they hold.

Task execution is the final layer in the system, and is managed by the Celery asynchronous task queue library [58]. This system acts as a light wrapper around tasks and can easily be swapped for alternative implementations. Using Redis as a message broker between the API and a task queue “worker”, tasks such as workflows are scheduled and then executed as resources are made available. In this manner, API tasks can be scaled horizontally, meaning more computer nodes running workers can be added, increasing the workload capacity for the system. Celery is primarily used for resource-intensive tasks such as workflow execution or monitoring.

2.2.3 Factory – Workflow Engine

“Factory” is a workflow execution system, that creates and runs pipelines through a schematic system. Given a remote/local repository of pipelines, it looks for YAML Ain’t Markup Language (YAML) [59] format files, to provide the API access to various workflows. On the client-side, available workflows are populated using the corresponding JSON schema via the API allowing new pipelines to be added dynamically. It is a “meta” system, as it is framework agnostic, meaning it can run any valid pipeline regardless of the implementation. Each schema consists of descriptor, resource location and command constructor elements (Figure 2.6). Descriptor elements provide metadata on the workflow itself, and its inputs. Workflows are composed according to the schema specification; workflows can be rigid and bound to a singular pipeline, or flexible, and combine multiple pipelines.

The primary driver behind Factory is containerization technology, where each workflow has a defined Docker image, hosted locally or in a cloud repository. Going outwards from the API to the web client, the workflow schema has predefined UI element types associated with input elements, which are used to construct forms. When a workflow has been submitted by a user via a web client, the associated workflow and input identifiers are then used to construct the workflow launch command. Using the Docker Python API library, and the Celery task execution system, the appropriate workflow image can be retrieved and then used to construct a container that will run that command.

Factory also generates reports for each workflow through IPYNB files co-located with each pipeline schematic in the provided repository. These files provide all the necessary information required to generate a workflow specific IPYNB, including data transformations, figure generation and descriptive text. In short, workflow outputs and any relevant tracking metadata are used here to generate an HTML report. The report is derived from the IPYNB file and a global HTML template, available to the entire system, ensuring that they are generated consistently.

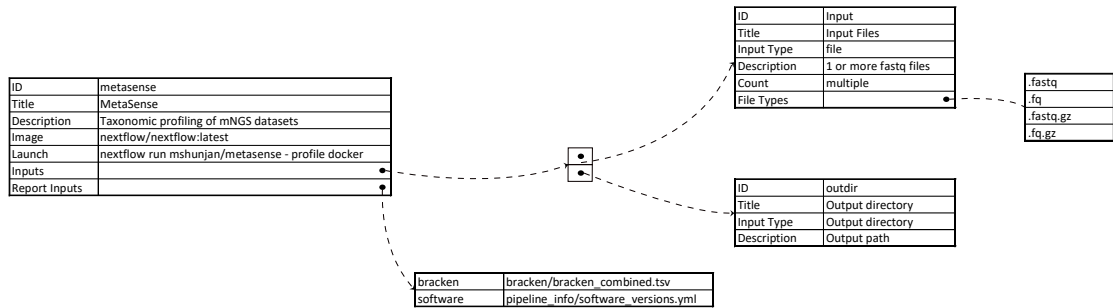


Figure 2.6 - Diagram of example Factory schema. Nodes indicate items, and arrows relationships between these items. Input items are categorized by the input type parameter.

2.3 Envisioned AlignDx User Workflow

The above-described components come together in a process that can generally be described in terms of the envisioned user workflow. First, a prospective user can access the web client through their preferred browser, and device using the website Uniform Resource Locator (URL). On arrival, they are greeted with some details on what the platform does, what it seeks to achieve, and next steps through a landing page. Access to web application services can be gained through registration. Once authenticated, users are redirected to a central dashboard, from which analyses can be submitted or later reviewed securely. From an “Analyze” tab, various workflows can be browsed, with associated metadata and descriptions. These include instructions on data requirements and expected outcomes of selecting the focused workflow. For example, current workflows include a data analysis pipeline for detecting a custom panel of viral pathogens in user-uploaded sequencing datasets (Chapter 3) and a line-detection pipeline for LFA images (Chapter 4).

Once a workflow has been chosen, the necessary data required for processing can be filled out via the prompted form and then submitted via a prompt button. In the event of a network issue, uploads can be resumed as necessary once connectivity is restored. On the server side, workflow processing is initiated, which initiates a recording cascade for execution, including inputs, outputs, and intermediary events. At this point, the workflow pipeline is on standby, however, once all input data has been uploaded, including any pending file uploads, execution begins. On the user side, both upload progress and analyses progress can be monitored via web client. Upon completion, workflow termination

triggers report generation, summarizing key findings using workflow outputs and tracking metadata, producing a final Hypertext Markup Language (HTML) report. This report, execution logs and output raw data are all stored on the server. Users can then access these reports in browser, and similarly download data. A record of these workflow submissions and corresponding results can be explored and manipulated from an “Archive” dashboard tab. This webpage is populated with this data alongside associated metadata, enabling more complex analyses by the user.

Chapter 3 – Genomics Workflows

Pathogen detection using genomic sequencing technologies presents a ripe opportunity for surveillance efforts. There are numerous sequencing-based approaches, with several different protocols and applications. In the domain of public health, key amongst them is whole-genome sequencing (WGS) of cultured isolates, and meta-omics (metagenomic or metatranscriptomic) sequencing of samples. With WGS, taxonomic classification is used to identify/verify the taxonomy of a single organism, but with meta-omics, bioinformatic methods are used to taxonomically profile a dataset. In surveillance applications with the latter approach, focusing on a subset of the resultant taxonomic composition becomes important. Specifically, this can be done by tailoring the analysis towards a “panel” of target pathogens and estimating their presence/absence or abundance.

Previously in Chapter 2, the AlignDx platform for running targeted bioinformatic workflows over the web was described. This chapter details the implementation of a genomic taxonomic classification workflow for human pathogenic virus surveillance via the AlignDx platform.

3.1 Base Pipeline

A general taxonomic classification workflow was developed using the Nextflow workflow system and the nf-core framework (Figure 3.1). Nextflow consists of a domain specific language (DSL), used to build bioinformatic pipelines [60]. This framework was chosen to demonstrate the “meta”-workflow capabilities of the AlignDx platform. Similarly to the Factory workflow engine, it takes advantage of containerization technology to bundle pipelines, aiming to provide reproducible, error-tolerant, and finally observable workflows [60]. The workflow also uses the nf-core framework, which automates pipeline creation and testing to standardize Nextflow implemented bioinformatic pipelines [61]. This gives it access to the repository of curated nf-core modules, making it flexible, and battle-tested. Additionally, it is tweakable, with the capability of restarting from the last common shared parameter thanks to the usage of Nextflow.

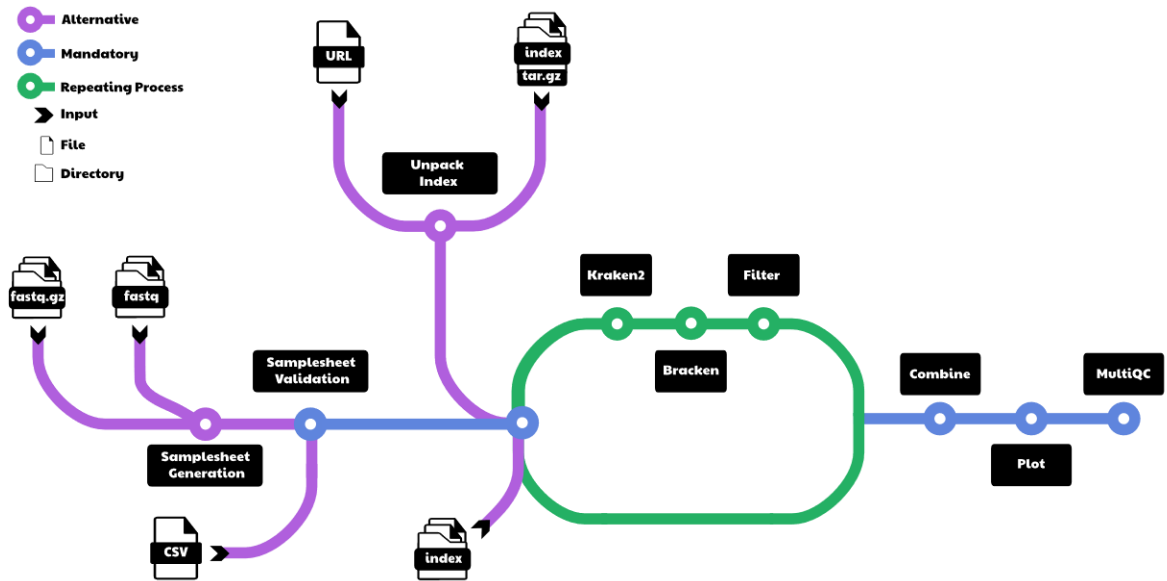


Figure 3.1 - Base pipeline architecture. Input genomic data is grouped according to samplesheet data, and then iteratively queried against a k-mer database until abundance measurements are collected and visualized. Round nodes indicate pipeline checkpoints, each making up separate nodes within the workflow. Inputs are all provided by the executor; either the user or system executing the pipeline. Adapted vectors and icons from nf-core pipeline components [29].

The pipeline is split by key processes that manage integral checkpoints that input data will flow through within an analysis. The first checkpoint evaluates the data input for correctness of type and format. FASTQ [62] or compressed FASTQ files act as input data types for this workflow and on encounter, the pipeline groups them according to metadata provided via a comma delimited file, generated automatically, or submitted by the user. Data can be grouped according to sample site or read pairing within this sample sheet. The submitted or automatically generated sample sheet is validated internally by the pipeline, and then files are made available for analysis. Optionally, files can be subsampled before taxonomic analysis using the Seqtk tool [63], either fractionally or using a fixed number of reads. This data is then taxonomically classified using Kraken2's [64] k-mer based approach with a chosen k-mer database. Briefly, k-mers are short genomic substrings, which Kraken uses with a reference database to classify reads via their lowest common ancestor taxa [64]. Kraken2 output files are then submitted to a Bracken [65] process for calculating the relative abundance of species within the sample. The resulting abundance measurements can then be filtered for contaminant reads, or any other organisms not of interest. Finally, abundance measurements for each submitted sample are collected and then used to generate interactive abundance plots via the Plotly visualization library [66]. All data output from the pipeline is additionally automatically aggregated into an HTML report via the MultiQC tool [67]. Workflow parameters, progress and errors are provided through tracebacks logged to a designated text file, alongside the executing terminal. Together, these components make up the base taxonomic classification pipeline and are used to construct genomic workflows used by the AlignDx Platform.

3.2 Audience Targeted Genomic Workflows

Using the base pipeline, three different genomic workflows were created, differing based on their output report targeted towards different audiences, defined within their pipeline schemas for the Factory workflow engine (Figure 3.2).

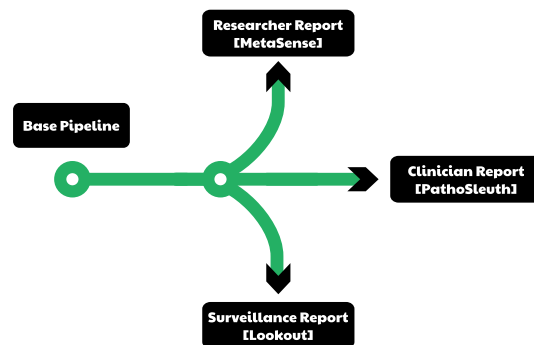


Figure 3.2 - Targeted genomics workflows. The depicted workflows continue from the last node of the base pipeline diagram (Figure 3.1). Nodes branch at the custom report, which summarizes the workflow data according to the target audience. Adapted graphics from the nf-core pipeline components [29].

For the MetaSense/researcher workflow, the report summarizes the unfiltered taxonomic profile of the submitted dataset. Using the bracken output results from the pipeline and the python Pandas data analysis library [68], the top 100 most represented species across each submitted sample are extrapolated. This is done by first extrapolating fractional abundance measurements from the bracken output of the base pipeline and using the DataFrame.nlargest function from Pandas to sort raw values in sample columns. It is difficult to show all species within the sample, as comprehensive Kraken2 k-mer databases, typically necessary for accurate taxonomic classification, can identify tens of thousands of organisms. Although this greatly depends on the nature and complexity of the sample, with metagenomics, a diverse taxonomic composition is often expected. Within a single HTML report, embedding larger datasets can reduce performance significantly, as it is not a data storage format. Additionally, many species may be taxonomically classified with negligible reads, thus a subset may still be representative of the taxonomic profile. The tabulated data is shown first in the report, and then additionally visualized in an interactive, relative abundance plot using the Plotly library in a stacked relative abundance bar plot. Finally, software versions used to analyze the input data and produce the

report are collated into a table, for transparency, which is done for each genomic workflow. Overall, the MetaSense report provides a glance at the taxonomic composition of the input dataset.

The PathoSleuth workflow is experimental, where it generates a binary detection report of pathogens within the submitted dataset. Similarly, to the MetaSense workflow, abundance measurements are extrapolated from the bracken output, but a subset of the raw table is used for further analysis, using a curated database referred to as the Pathogen Panels DB, described in the next section. The fractional measurements are then converted to a Yes/No classification for each sample, based on an abundance threshold value that can be defined in the system (default value = 0). Finally, the Lookout/surveillance workflow similarly uses this DB but instead provides raw detected read numbers, fractional abundance measurements and human filtered abundance measurements.

3.3 Pathogen Panels DB

The Pathogen Panels Db is an ongoing flexible flat-file database that detects 228 human pathogen viruses (Figure 3.3). Viruses are the focus of this database, as bacterial pathogens introduce additional complexities to taxonomic classification in humans. Organisms are listed by genus and species, alongside their corresponding National Center for Biotechnology Information (NCBI) taxonomy ID. Organisms are classified by a domain of interest using a “panel”, through a binary classification. Currently, three panels are in use: a COVID-19 panel for detecting SARS-CoV-2 virus, a CDC high-consequence virus panel and finally a Human Pathogenic Viruses panel. The CDC high-consequence panel is tailored according to viruses listed by the Division of High-Consequence Pathogens and Pathology (DHCPP) [68]. Finally, the Human Pathogenic Viruses panel makes up the entirety of the database. This panel was built as a team effort (M. Hunjan, N. Abu Mazen, A. Doxey) by curating human pathogenic viruses based on several sources (CDC, Health Canada, Exspasy ViralZone, and others).

Organism	TaxID	Type	START	COVID-19	Human Pathogen	Human Pathogenic Viruses	CDC high-consequence viruses
Adeno-associated virus		272636 Virus		N	Y	Y	N
Aichi virus 1		1313215 Virus		N	Y	Y	N
Andes orthohantavirus		1980456 Virus		N	Y	Y	Y
Asama orthohantavirus		1980457 Virus		N	Y	Y	Y
Asikkala orthohantavirus		1980458 Virus		N	Y	Y	Y
Australian bat lyssavirus human/AUS/1998		446562 Virus		N	Y	Y	N
Banna virus		77763 Virus		N	Y	Y	N
Barmah forest virus		11020 Virus		N	Y	Y	N
Bayou orthohantavirus		1980459 Virus		N	Y	Y	Y
Black Creek Canal orthohantavirus		1980460 Virus		N	Y	Y	Y
Black Creek Canal virus		1980460 Virus		N	Y	Y	N
Bombali ebolavirus		2010960 Virus		N	Y	Y	Y

Figure 3.3 - Snippet of the Pathogen Panels Db.

3.4 Wastewater – Surveillance Workflow Run

With continued interest in making a surveillance platform, further experiments were done with the Lookout/surveillance workflow. To demonstrate its utility using the AlignDx platform, a wastewater dataset was selected and run through the system (Figure 3.4). The dataset is 0.15 GB in total size, consisting of paired-end compressed FASTQ format sequence data generated from two different RNA shotgun sequencing runs on wastewater treatment plant samples obtained from collaborators (Charles Lab, U. Waterloo). These were 24-hour composite samples taken by autosamplers from the following plants: GE Booth Lakeview wastewater treatment plant, the York-Peel OCF/Humber wastewater treatment plant, the Kitchener wastewater treatment plant and the Waterloo wastewater treatment plant.

First, a user was registered with the platform, and after signing in with credentials, and navigating to the dashboard, the Lookout/surveillance workflow was chosen using the pipeline select menu (Figure 3.4A). The input form elements (Figure 3.4B) were dynamically generated after selection, alongside metadata on the workflow, and its inputs. These were then filled out using the run name “wastewater”, the pathogen panel “Human Pathogenic Viruses” and all input FASTQ files from the dataset. Submission progress was monitored using the status card and upload progress bars in the monitor UI (Figure 3.4C). Once the status card indicated that the run had been completed, the report could then be seen (Figure 3.4D) and then later reviewed in the archive (Figure 3.4E), when necessary. From the report, two of the samples were found to contain reads matching SARS-related coronavirus, and no other viral pathogens. Using the archive, this report alongside the raw Bracken and Kraken2 results and metadata on submission time, status, and duration could be explored.

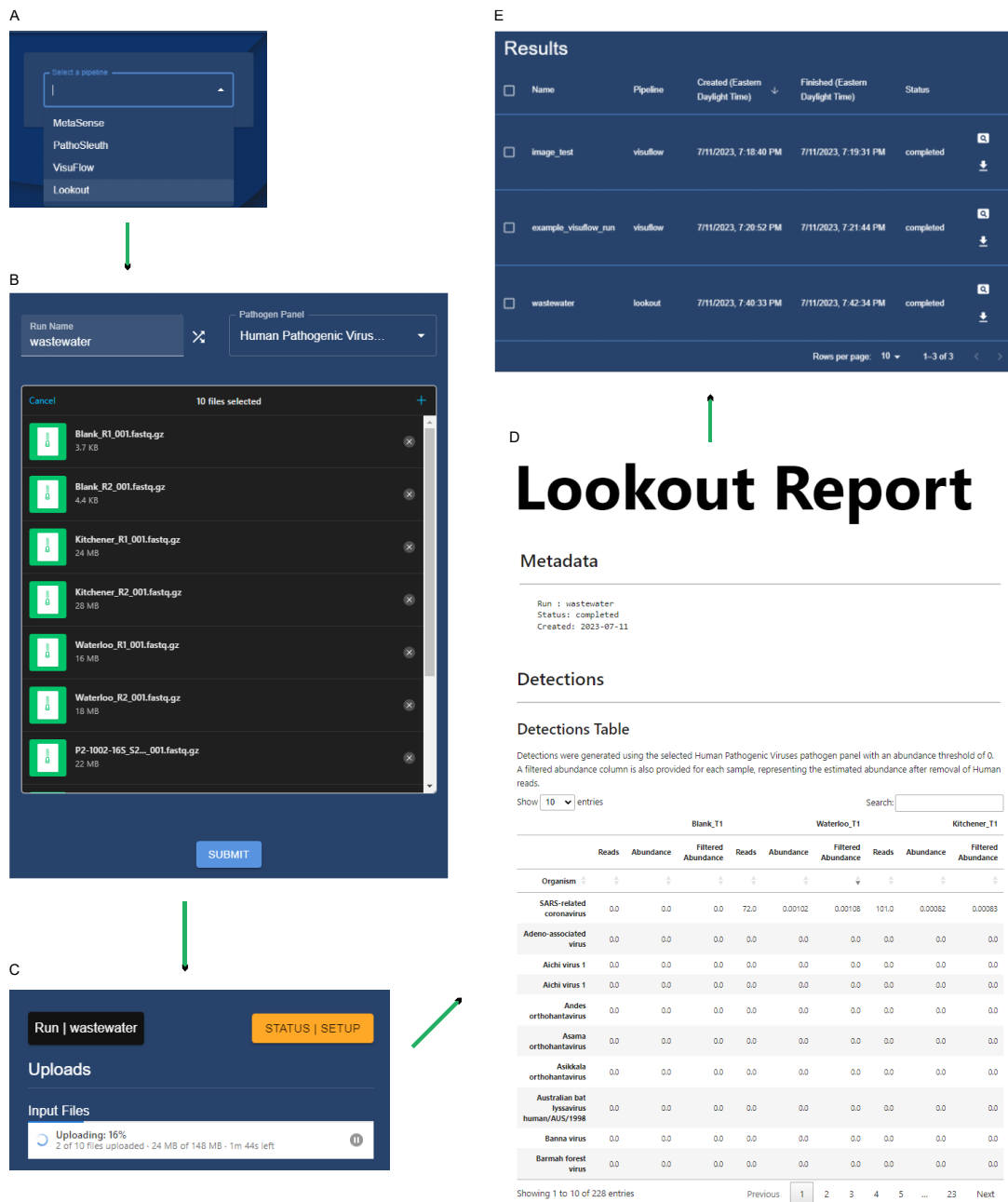


Figure 3.4 - Surveillance/Lookout workflow via AlignDx client UI. Stepwise utilization of Lookout in the AlignDx Platform. (A) Workflow/pipeline selection menu and available workflow options. (B) Filled out surveillance form, with input dataset, and chosen pathogen panel. (C) Status card for monitoring submissions. (D) Wastewater dataset report. (E) Archive entry for submission. Adapted vectors from nf-core pipeline components [29].

3.5 COVID-19 clinical swab dataset

With an interest in the performance capabilities of the base pipeline, a clinical dataset was chosen for testing. Through collaborators (Dr. Samira Mubareka, Sunnybrook Health Sciences Centre), a shotgun RNA-seq dataset was obtained of nasopharyngeal swabs collected from 66 patients throughout October 2020 – 2021. Samples underwent qPCR analysis for SARS-CoV-2 and were then labelled positive and negative accordingly based on standard cycle thresholds (Ct). Samples were also split into outpatient, ICU or Non-ICU categories. Samples were sent for sequencing in November 2021. RNA-seq libraries were prepared from 50ng of RNA samples via the NEBNext rRNA Depletion Kit v2 (Human/Mouse/Rat) (NEB Cat# E7405) in conjunction with NEBNext Ultra II RNA Library Prep Kit for Illumina (NEB Cat# E7765). rRNA depleted libraries were pooled equimolar and quantified using the NEBNext Library Quant Kit for Illumina (NEB Cat# E7630). Finally, libraries were sequenced 151c paired-end at 65M per sample on the NovaSeq6000 platform at the Donnelly Sequencing Centre (Toronto,ON), using the S2 v1.5, 300-cycle kit (Illumina Cat# 20028314). The sequencing reads displayed a high percentage (~85%) of uniquely mapped reads, and the correlation between sequenced replicates was high (Pearson Correlation $R \approx 0.9$). A total of 66 paired-end read FASTQ files, making up 667 GB of data, was generated by this process in three batches.

The base pipeline was then run on this dataset for each data batch, using the publicly hosted Standard Kraken2 DB capped at 8 GB, run with Bracken at a species level resolution, filtering out human host reads. The combined Bracken output fractional data was then used to visualize the taxonomic profile of the data using a clustered heatmap, where SARS-CoV-2 was found to cluster with high fractional abundance on positively labelled samples (Figure A.1). For further characterization of the dataset, SARS-CoV-2 abundance measurements were extrapolated from the collated data to overview relative abundance across samples (Figure 3.5). Positively identified samples using the base pipeline were found to corroborate qPCR labels from provided metadata.

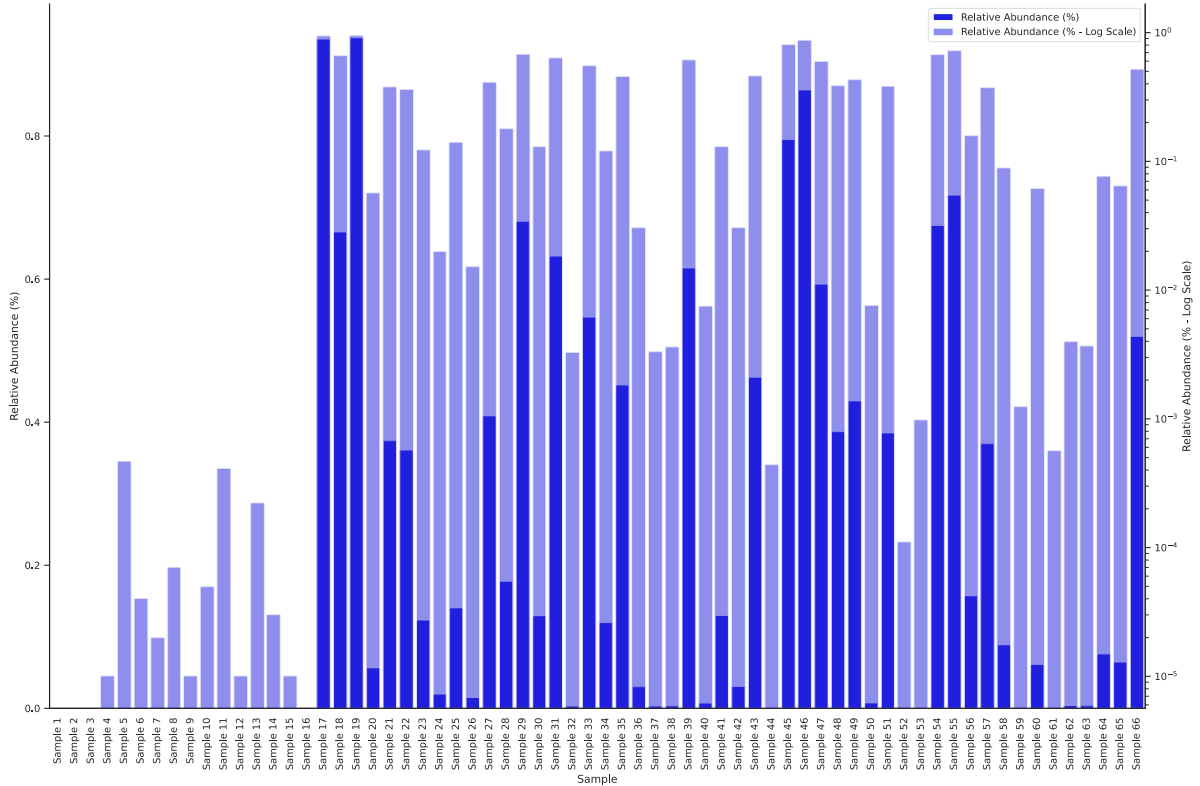


Figure 3.5 - Proportional abundance of SARS-CoV-2 across samples. Clinical swab samples were analyzed using the base pipeline, producing a combined Bracken output containing the taxonomic profile of the dataset. SARS-CoV-2 abundance values were extrapolated from these results and then graphed in a bar plot. Dark blue bars represent proportional abundance on a 0-1 scale, and light blue bars represent the log-scale representation of these values. Samples are ordered as follows: 1-16 (Negative), 17-32 (Outpatient), 33-48 (non-ICU), 49-66 (ICU).

3.6 Performance Evaluation

To get a sense of the sequencing depth required to make accurate predictions to the presence/absence of SARS-CoV-2 in this dataset, various performance metrics were calculated. Using qPCR metadata as “ground truth” labels for samples, the same analysis from above was re-run, but with fractional and fixed read reservoir subsampling at magnitudes of 10. Fractional subsampling was performed for values 10%, 1%, 0.1%, 0.01%. Fixed read subsampling was performed for values 10M, 1M, 100K, 10K, 1K, which were chosen based upon the smallest file size in all batches, being ~44 million reads. First, an array of thresholds was generated for each sampling condition based on the relative abundance scores for each sample. Each of these thresholds were then used to classify samples as SARS-CoV-2 positive or negative, using binary values of 1 and 0 respectively (Figure 3.6).

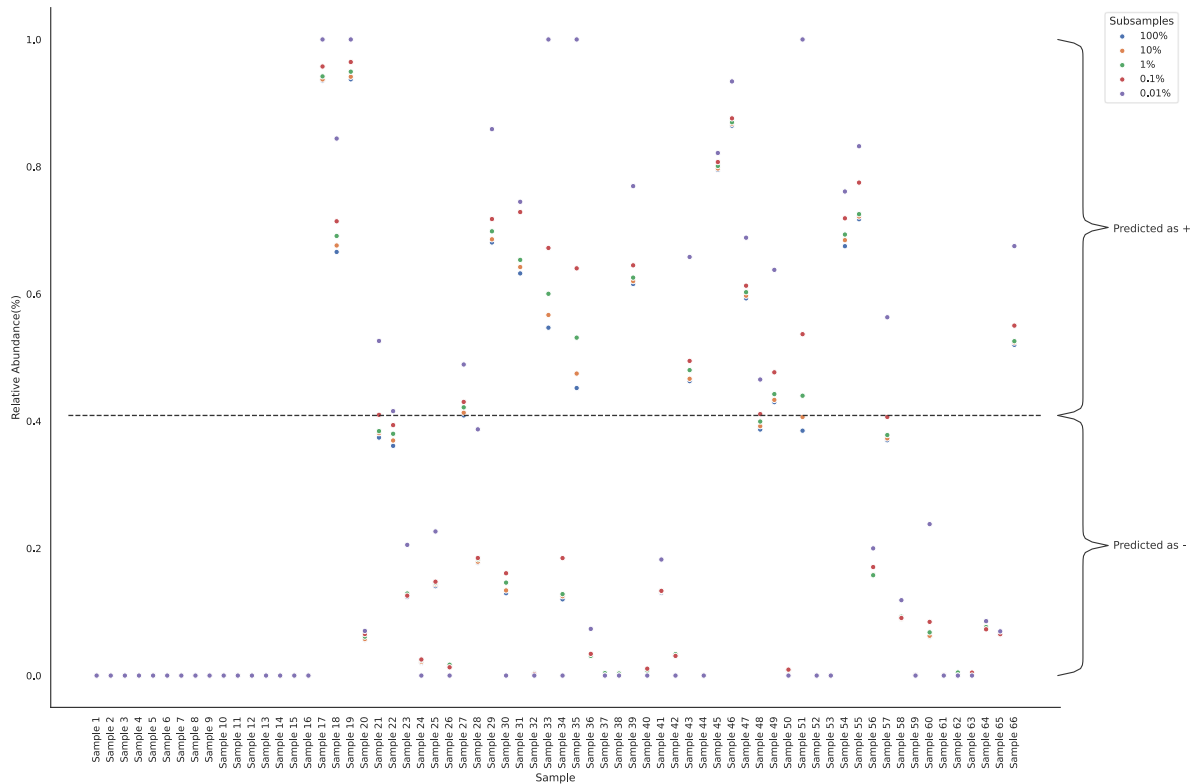


Figure 3.6 - Example scatterplot for SARS-CoV-2 detection classification. Clinical swab samples were separated using a threshold of 0.40892 for classification of SARS-CoV-2. Data points above this threshold were predicted as positive, and below as negative. Samples are ordered as follows: 1-16 (Negative), 17-32 (Outpatient), 33-48 (non-ICU), 49-66 (ICU).

These detection predictions, alongside the positive and negative labels from the SARS-CoV-2 dataset, were manipulated using Pandas, and then used to generate confusion matrices with the Sklearn `confusion_matrix` function. True negative, false positive, false negative and true positive values were retrieved from the flattened matrix and used to calculate multiple metrics for performance evaluation at each threshold for unsampled (Table A.1), fractional (Table A.2-5) and fixed subsampling (Table A.6-10). At a threshold of 0.0056, for reads required to classify a sample as SARS-CoV-2 positive, accuracy was found to be 97 % (Table A.1). The calculated TPR and FPR values were then used to visualize the performance of the base pipeline for the subsampling conditions, as well as unsampled data, using Receiver Operating Characteristic (ROC) curves (Figure 3.7A, B). The pipeline was found to classify the clinical swab dataset reads according to the qPCR metadata very well across all subsampling conditions, until subsampling conditions reach below 0.1% or 100k reads. The trend in the corresponding Area under the ROC Curve (AUC) scores was also visualized for each of the subsampling conditions (Figure 3.7C, D). This similarly reflects the observations seen in the ROC curves; as subsampling conditions get deeper; the pipeline performance gets worse.

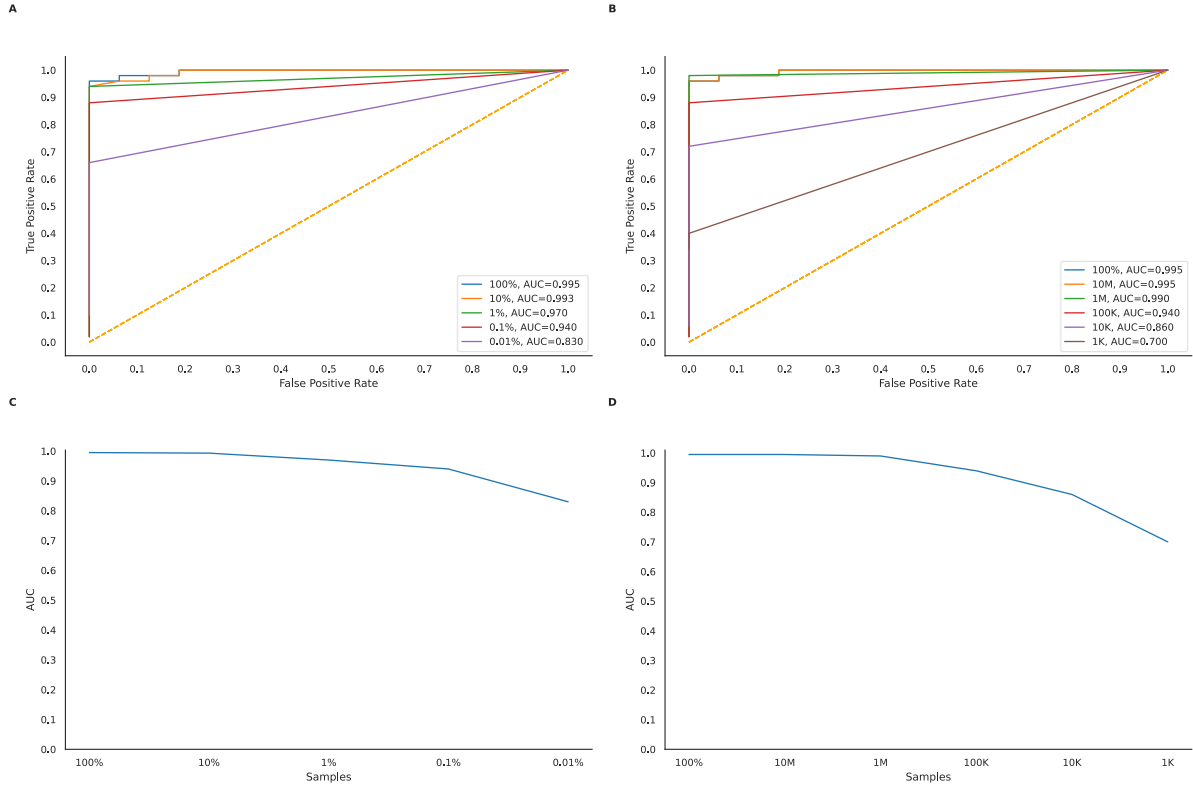


Figure 3.7 - Accuracy of SARS-CoV-2 detection on the clinical swab RNA-seq dataset using the AlignDx genomic workflow. (A) ROC curves across multiple thresholds for fractional sampling conditions. (B) ROC curve across multiple thresholds for fixed read sampling conditions. (C) AUC curve across fractional sampling conditions. (D) AUC curves across fixed sampling conditions.

Chapter 4 – LFA Workflow

Pathogen detection approaches such as LFA tests have been critical during the COVID-19 pandemic towards surveillance efforts. As a POC approach, they are key in public health, offering low-cost, portability and ease-of-use. Alongside these factors, the maturity of this technology has been an important factor in creating this opportunity. Together, these strengths are realized by both healthcare professionals and everyday individuals to improve public health efforts. Although this enables decentralization, with the plethora of LFA test kits available commercially, and a lack of standardization industry wide, reliability of test interpretation can be variable. An integrated approach leveraging smartphone camera technology, along with the security, UI experience and archiving capability of AlignDx can provide consistency. This can further be adopted in healthcare settings to monitor outbreaks through mass-surveillance. Thus, in this chapter, a POC approach to mass automated LFA-based pathogen surveillance via the AlignDx platform is described.

4.1 LFA Tests

Three commercially available LFA tests from different suppliers were purchased to generate a training image dataset for an LFA diagnostic machine learning algorithm “VisuFlow”, representative of a range of potential real-world conditions. The three tests purchased were, i) Quidel Strep A, ii), ICON DS Strep A, and iii) BD Veritor Flu A+B. An undiluted positive control sample was generated according to manufacturer directions, to act as a maximum signal for each test in the dataset. These samples then underwent a serial half dilution (undiluted to 1/64) with a diluent control representing a negative condition (Figure 4.1).



Figure 4.1 - Overview of serially diluted LFA tests under controlled light conditions. Representative images of the three commercially available (A: Quidel Strep A, B: ICON DS Strep A, C: BD Veritor Flu A+B) LFAs performed in serial half dilutions from 1/2 to 1/64 using the provided positive control reagent under 5500K light. D: Representative images of the controlled lighting conditions that each assay and dilution series were collected under (The Quidel test is shown as example). Lighting conditions started at 3200K (upper left image: warm – orange tint) and underwent twenty-four 100K increments to 5500K (bottom right image: cold – blue tint).

The ICON DS and Quidel tests were performed according to manufacturer’s direction with results read “wet” immediately following 5 minutes of LFA being immersed in sample. Additional images were taken for the ICON DS and Quidel tests 30 minutes following test completion to represent a “dry” condition that would occur if a user did not analyze their assay as directed by the manufacturer. The “wet” and “dry” conditions were selected to represent the ideal (wet) and sub-optimal (dry) conditions where a LFA test would be analyzed. The BD Veritor tests were removed from the OEM casing to enable image acquisition under controlled lighting in a narrow field of view, resulting only in “dry” images being collected. Ground truth labels were assigned to each test for each condition. Furthermore, undiluted samples for BD Veritor had smeared control lines, making it challenging to assign a ground truth label to this subset of the data, therefore it was excluded from further analysis.

4.2 Training Dataset

LFA images were captured under controlled lighting conditions using a Canon 5DMK3 full frame digital single lens reflex camera with 35mm Canon L lens, aperture f4.0, shutter 1/200, and file format as CR2 RAW (Figure 4.2).

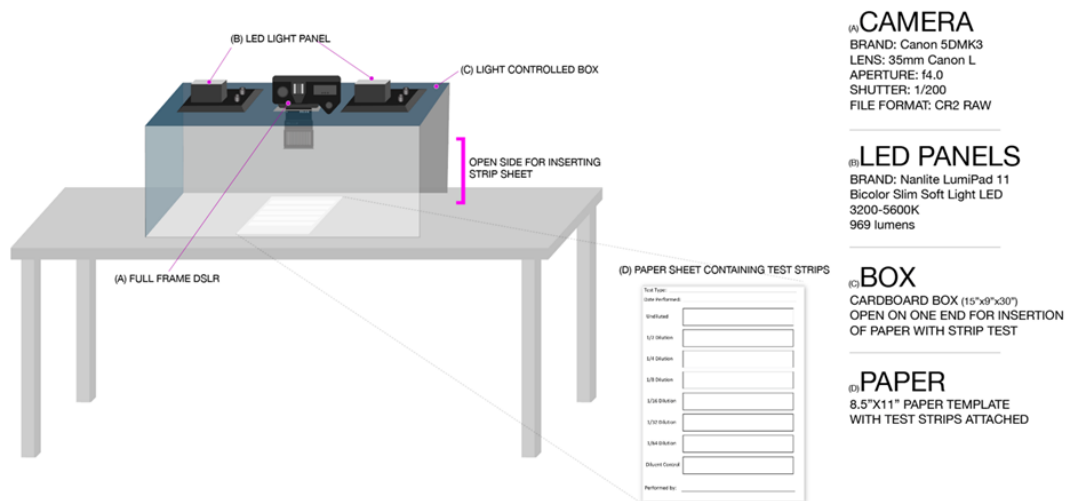


Figure 4.2 - Schematic of image acquisition workflow.

Lighting was provided by two Nanlite LumiPad 11 BiColor Slim Soft Light LEDs placed in a custom light box that would exclude ambient room light. The lightpads allowed for 100 kelvin light temperature ramps and image acquisition from 3200 kelvin (warm) to 5500 kelvin (cool). This was done to model real-world varying light conditions that operators would face when analyzing LFA tests. RAW (DNG) image formats were then converted to JPG images for easier downstream analysis. Using this process, 24 images were collected from the three serially diluted (8 total dilutions each) LFA tests resulting in a training image dataset of 960 images.

4.3 The VisuFlow Pipeline

An LFA based pathogen detection pipeline “VisuFlow” was developed in python using OpenCV (Figure 4.3). First, the training dataset was used to establish optimal line detection, noise identification, and result output. JPG images from the training dataset were first cropped manually to simulate user operations. Then images were further cropped programmatically using a custom function based on the OpenCV and Numpy python libraries. This reduced image border by a fixed value of 10%, focusing detection on the LFA and removing extraneous noise from the dataset. Additional image smoothing and noise-reduction was done using a gaussian filter (block size = 5x5 pixels). Finally, images were resized to fixed dimensions (512 height x 256 width pixels).

Local adaptive thresholding was done using OpenCV to binarize the image using a block size of 2.5. This step assigns pixels to a binary value according to an intensity threshold. The threshold is determined based on local regions within a single image. Pixels exceeding the threshold are assigned a value of 1 and below 0. Following binarization, further noise was eliminated, alongside smoothing using OpenCV morphological operations to generate a 10x10 pixel kernel. Finally, a sliding window (width = image width, height = 10 pixels) was applied from the top to the bottom of the image. For each window, a fraction was measured based on pixels assigned a value of 1. Using a fraction threshold of 0.8, windows above were predicted as a positive window (contain an LFA line segment) and two consecutive positive windows identified a complete LFA control/test line. The algorithm would repeat this method to find additional lines, and then the algorithm would output the number of lines detected, and optionally, an image displaying the detected lines overlaid on the original image.

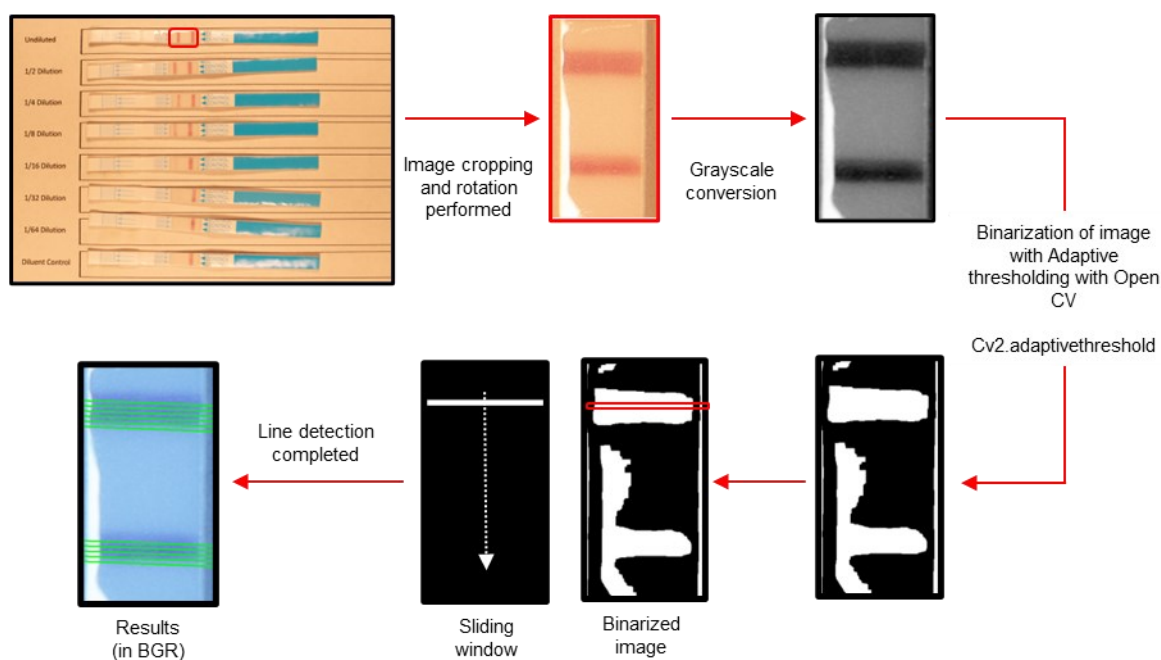


Figure 4.3 - Overview of the Visuflow pipeline. Lateral flow assay images were cropped to reveal only the region of interest for analysis, followed by grey scale conversion, binarization of the image using OpenCV adaptive threshold function (see methods), and line detection using a slide window function. Results are displayed in RBG color.

4.4 Performance and Validation

A metric analysis of the pipeline was performed by comparing its predictions to the manually labeled images (considered “ground truth”). These values were manipulated using Pandas, where labels were binarized as 1 for positive, and 0 for negative values. Confusion matrices were then generated for each block size of each test using the Sklearn `confusion_matrix` function. True negative, false positive, false negative and true positive values were retrieved from the flattened matrix and used to calculate multiple metrics (Appendix A - Calculations) for performance evaluation (Table A.11-14). Predictions were as follows:

True positive (TP) tests were those for which two lines were correctly predicted. True negative (TN) tests had one-line (positive control) lines correctly detected. False positives (FP) and false negatives (FN) were quantified in parallel for downstream use in sensitivity, specificity, and F1 score calculations. Finally, two exception cases were handled: if zero or more than two lines were detected, it was flagged as an error.

It was found that the size of the pixel neighborhood used to calculate the pixel threshold (block size) impacted image binarization greatly (Figure 4.4). All 960 images were analyzed under 2-unit incremental increases in the hyperparameter (41 through 281) for all metrics.

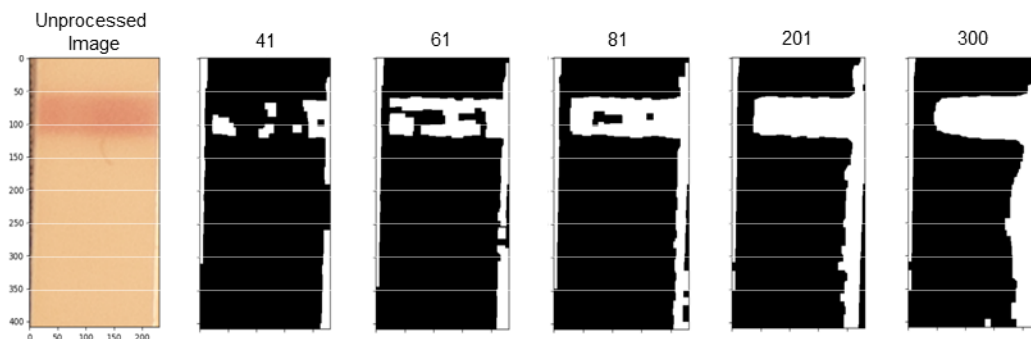


Figure 4.4 - Impact of hyperparameter variation on image binarization. Hyperparameter variation impacts binarization of image and contributes to variation in line detection. Representative binarized images under increasing hyperparameter values for the ICON Strep A lateral flow assay are shown.

ROC curves were then used to visualize performance (Figure 4.5). Across multiple hyperparameters (block size, light temperature, and dilution), the pipeline performed robustly on this dataset, with little variation in AUC scores. Although performance was relatively stable across the tested hyperparameters, further exploration was done with the block size and light temperature parameters.

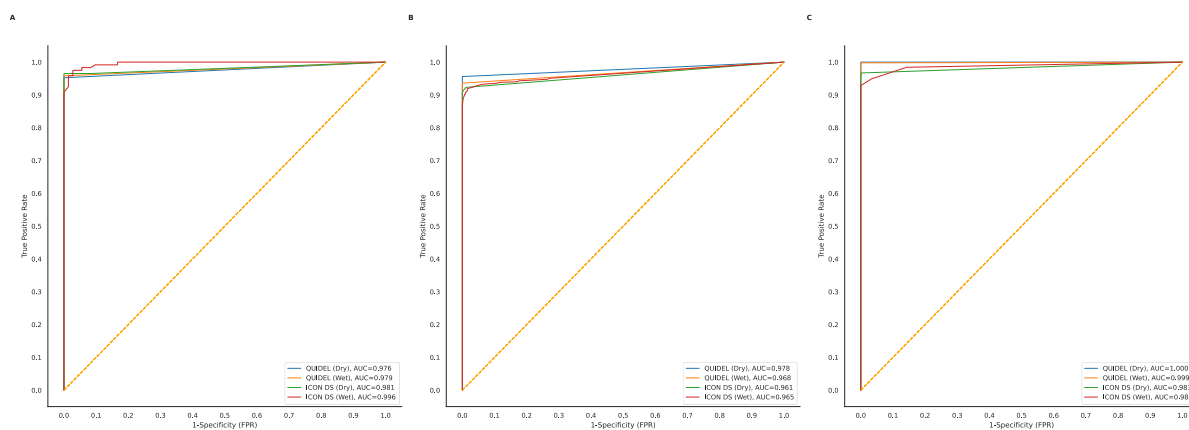


Figure 4.5 - ROC Curve across multiple hyperparameters for each test. AUC scores were calculated using the Sklearn AUC function. (A) ROC based on block size variation. (B) ROC based on temperature variation. (C) ROC based on dilution variation.

Analysis was performed with and without the BD Veritor LFA images as the assay presented the least visible lines under even ideal light conditions and in non-diluted samples. Across all LFA tests including or excluding the BD Veritor images, the hyperparameter set at 201 generated the largest F1 score (0.944 with BD Veritor, 0.977 without BD Veritor) and was chosen as the optimal value. The average TP rate and FP rate for all LFA tests over a range of hyperparameter values were used to generate a ROC curve (Figure 4.6). Performance improved with the exclusion of the BD Veritor images.

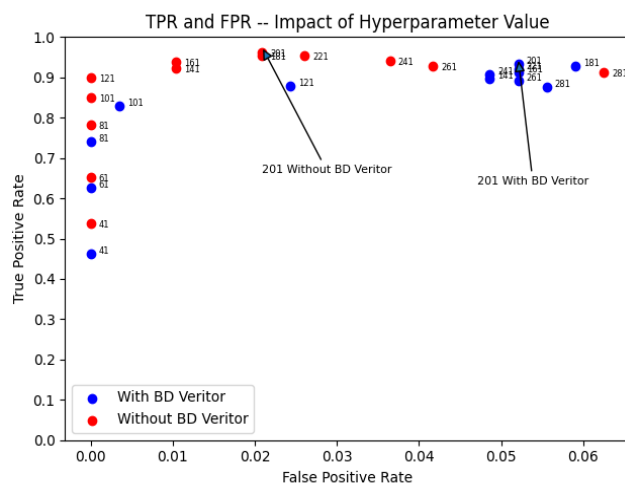


Figure 4.6 - Impact of hyperparameter value with and without BD Veritor. Model performance was evaluated using TPR and FPR performance metrics, with removal or inclusion of BD Veritor data points. The hyperparameter values for each datapoint are highlighted.

Using the optimized hyperparameter value and the ground truth labelled images, the impact of lighting temperature on TP rate, FP rate, and F1 score were next investigated (Figure 4.7). Irrespective of the inclusion or exclusion of the BD Veritor LFA images, it was found that 100 kelvin lighting ramps from 3200 kelvin (warm) to 5500 kelvin (cool) had minimal impact on the TP rate, FP rate, and F1 score metrics of performance.

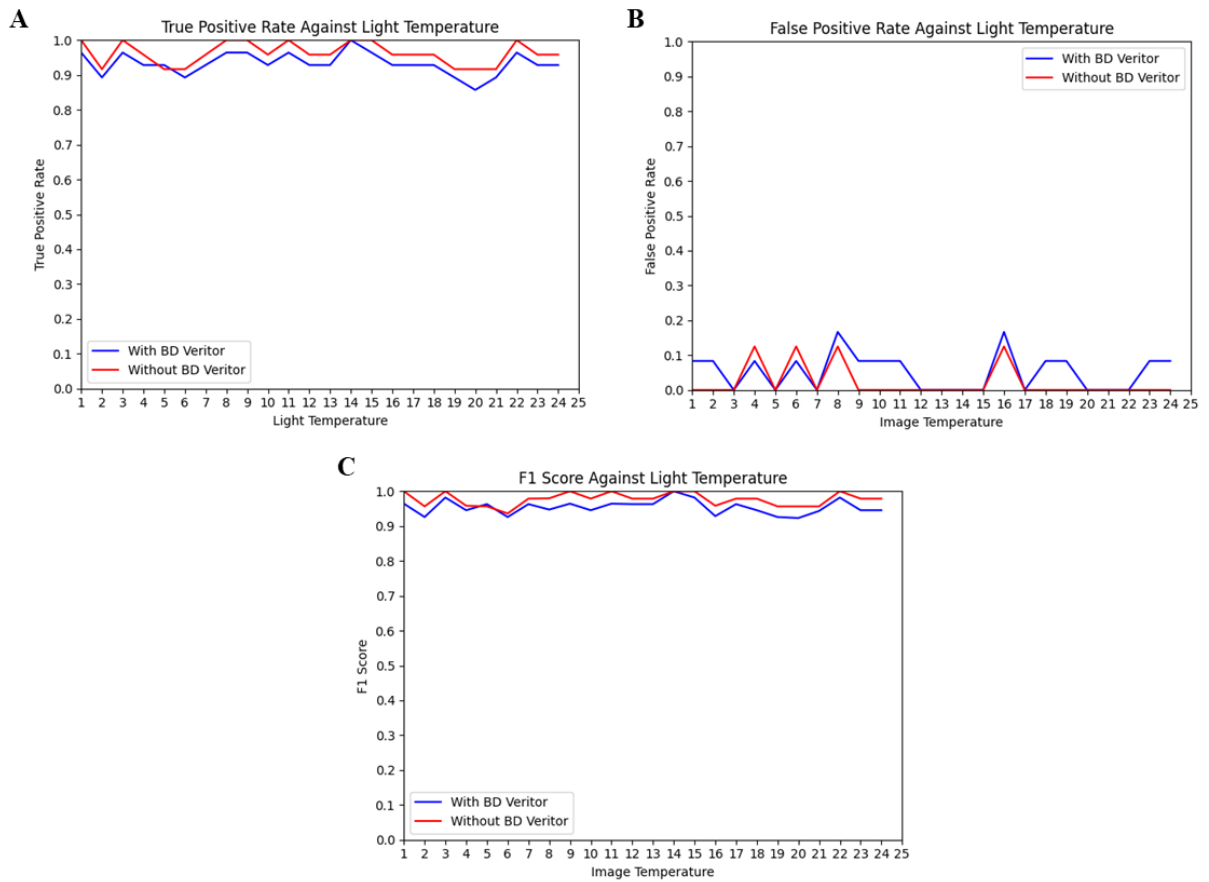
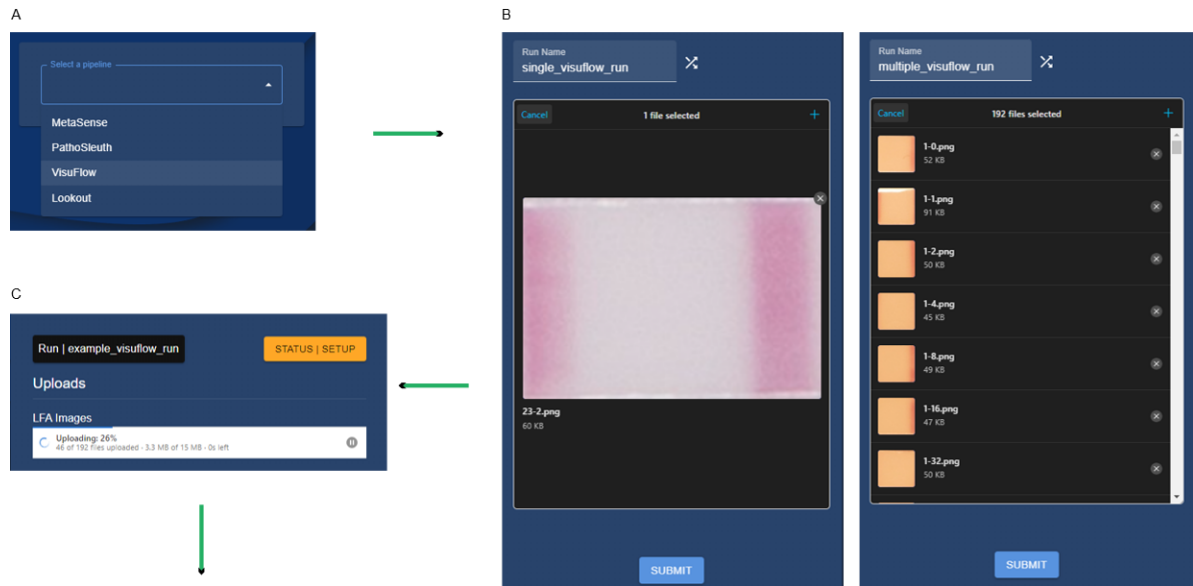


Figure 4.7 - Impact of lighting temperature on image analysis algorithm performance. TPR (A), FPR (C) and F1 (E) statistics were calculated for all five lateral flow assays under a range of colour temperatures including (blue) or excluding (red) the BD Veritor data.

4.5 Implementation of VisuFlow in The AlignDx Platform

To demonstrate the utility of the VisuFlow workflow using the AlignDx platform, the training dataset from above was selected and run through the system (Figure 4.8). User credentials were authenticated via the client sign in form, redirecting to the central dashboard, where the VisuFlow pipeline could be selected (Figure 4.8A). Once selected, the input form components were generated, consisting of a run name input and an image upload input. Images could additionally be taken using a native camera on mobile/desktop devices or uploaded from remote cloud storage services. Two analysis paths were chosen to demonstrate the capabilities of this system: a singular image analysis (Figure 4.8B - Left) and multiple image analysis (Figure 4.8B - Right). Singular image analysis was performed to represent a self-testing scenario, where an individual at-home could perform an LFA test according to the manufacturer's direction, and then utilize this workflow to determine a detection result. Multiple image analysis falls in line with mass surveillance, which could be utilized for outbreak monitoring, amongst other surveillance purposes. The analysis progress was monitored using the status card UI (Figure 4.8C) until all images were uploaded, at which point the VisuFlow report could be viewed (Figure 4.8D). Finally, metadata regarding the submission was reviewed in the archive (Figure 4.8E), alongside downloading raw output data from the workflow.



D Visuflow Report

Metadata

Run : example_visuflow_run
 Status: completed
 Created: 2023-07-11

Detection results

Show 10 entries Search:

	Sample	Detected Lines	Result
0	9-0.png	1	negative
1	4-32.png	2	positive
2	7-32.png	2	positive
3	14-64.png	1	negative
4	12-32.png	2	positive
5	19-32.png	2	positive
6	13-32.png	2	positive
7	4-64.png	1	negative
8	16-8.png	2	positive
9	17-2.png	2	positive

Showing 1 to 10 of 192 entries Previous 1 2 3 4 5 ... 20 Next

E Results

<input type="checkbox"/>	Name	Pipeline	Created (Eastern Daylight Time)	Finished (Eastern Daylight Time)	Status	
<input type="checkbox"/>	image_test	visuflow	7/11/2023, 7:18:40 PM	7/11/2023, 7:19:31 PM	completed	
<input type="checkbox"/>	example_visuflow_run	visuflow	7/11/2023, 7:20:52 PM	7/11/2023, 7:21:44 PM	completed	

Rows per page: 10 1-2 of 2

Figure 4.8 - Visuflow workflow via AlignDx client UI. Stepwise utilization of Visuflow in the AlignDx Platform for single/multiple LFA image analysis. (A) Available workflow/pipeline workflow options. (B) *Left* – Single image upload; *Right* – Multiple image upload. (C) Status monitor. (D) Image analysis Report. (E) Archive entry for submission. Adapted vectors from the nf-core pipeline components [29].

Chapter 5 – Discussion and Conclusions

5.1 The AlignDx Platform

5.1.1 Challenges in Digital Surveillance

The pandemic raised alarms regarding the current state of the public health sector, calling for the improvement of public health countermeasures, specifically mitigation strategies such as surveillance. Additionally, with the increasing trends in EIDs [8], surveillance is an integral barrier to public safety. From the context of surveillance data inputs to the outgoing reports, gaps appear in the diversity of data sources and the intended target audiences of reports. The AlignDx platform seeks to solve these limitations, ultimately bettering these systems, and the corresponding public health response.

Surveillance data is typically tightly coupled into the premise of the surveillance platform, making flexibility in data sources across domains quite difficult. Nextstrain for example, heavily focuses on strain-level epidemiology using publicly available sequencing data, and its associated metadata [40]. Consider the study by Karr et al., [69], which identified 26963 potential coronavirus genomes from the NCBI, with questionable data quality. Curating this data case by case becomes essential and is not feasible at scale without additional processing steps. Bias is inevitable as well, with resource inequity, meaning the dataset may not be representative of the larger population. In the case of cost, although sequencing has gotten cheaper, the availability of sequencing instruments alongside reagents may be limited in low-middle income countries, making global surveillance difficult, via this platform. The interdisciplinary nature of public health necessitates a general multi-domain approach, where any data source, and subsequent workflow can be incorporated into the platform. Data-driven systems of the modern era are strengthened by their variety, and can provide necessary contextualization for downstream decision making [70]. AlignDx attempts to solve this issue by supporting any workflow, no matter the domain, through the Factory workflow engine. Data sources do not just have to be genomic, nor do they have to be of a singular type; a variety of input streams can be utilized to perform surveillance analyses. Data sources are tied to the workflow, so they can be as simple, or as complex as desired, even replicating analyses and reports produced by other platforms, such as those provided by Nextstrain. Of course with varying sources, issues of data integration and interoperability begin to arise [70]. In a review of a variety of technologies for surveillance in Tanzania by Mustafa and colleagues [71], researchers found fragmentation and a lack of interoperability, limiting the impact of otherwise promising surveillance endeavors. The AlignDx platform includes these considerations

through separation of concerns, where each major component makes up a module of the larger system. In the case of incoming data, interoperability becomes increasingly difficult, especially as data variety increases. As workflows are independent of each other, each phase of the pipeline, including input processing, is defined only by that workflow, and its corresponding web UI form.

Data privacy is another large concern with surveillance systems, as the abundance of sensitive health data flowing through them brings up legal and ethical challenges. Health-relevant data has typically been a challenge to integrate in data science, due to complex security regulations [72]. By giving users full control over their data, AlignDx can circumvent many of these issues through its privacy first design. Bentotahewa and colleagues suggest that only required information should be collected by the system, following data minimization principles [73]. Crafting guidelines for mandatory and supplemental information is subjective but aims to follow the demands of the system. In the example of user registration, only authentication-relevant information is stored within the system database. Broadening to the perspective of workflows, all submitted form data is stored, but anonymized. These are, notably, accessible for review through the archiving system, where they can be removed as desired. Another key aspect of data minimization is limited data retention, where user submitted information is only retained for a limited time frame [73]. In the case of uploaded data, such as genomes, AlignDx prunes them immediately upon completion. Besides the cost and performance benefits of removing this data from the hosting servers, this greatly reduces the risk of any potential data breach. Although workflow outputs and reports are stored by the platform, this data is anonymized and can only be linked to a user through database queries.

The target audience of surveillance platforms can also be quite narrow, which can act as a barrier to the public health response. In terms of outgoing responses, this can be solved through a standardized reporting system, tied directly to each workflow. Reporting in the modern surveillance pipeline can be thought of as a series of steps beginning with researchers and ending with the public health response [74]. AlignDx provides such a system where reports can be tailored to any entity within that reporting hierarchy, as well as reviewed at any time. Reports are disseminated via the web using the HTML format, and thus can be viewed through the web platform, or locally through a browser on any device. Workflow outputs are additionally accessible and can be used by other platforms or tools as desired, playing into outgoing interoperability. UI is a less-considered challenge in the realm of surveillance platforms; few, if any bioinformatic tooling, alongside platforms, are intuitive to use. Some research has explored the impact of UI in the context of disease surveillance systems, suggesting a

simplified/minimal UI that focuses on dynamic elements [75]. A myriad of options and layers provide researchers excellent control over inquiries but act as a barrier to interpretation. On the opposite extreme, purely functional design can be prohibitive in a web medium that is privy to intuitive design patterns. AlignDx therefore strikes a balance on this spectrum, with a minimal, consistent, but flexible UI design.

5.1.2 Design advantages of AlignDx

The modular design of AlignDx through its container-based service architecture is key to the longevity it can sustain as a surveillance platform. Each service communicates primarily through common networking protocols, meaning that maintenance, or replacement is service dependent. This brings with it the capacity to keep up with rapid technological changes. One study on technological improvement rate in the United States found that the majority of fast improving technologies, being those with greater than 36.5 % improvement per year, were dependent on software [76]. Turnover rates in software longevity specific to web development have not been comprehensively studied, but the same study identified web-specific domains such as networking, encryption, data flow and software delivery methods as among the 20 fastest improving technological domains [76]. In the span of this thesis alone, authentication, UI, and workflow technologies in the AlignDx platform have been superseded by novel technologies, which have been used in replacement. The flexibility of this design pattern means architecture is not tied directly to the software implementation, at the cost of increased required expertise. Better service implementations can be prioritized, leading to a more efficient platform, overall. Coordinating these services and managing them as separate entities is largely a result of containerization. Containers acts as standard units of software that can be scaled, reproduced and managed [77]. This standardization provides a necessary layer of control over the various services offered by AlignDx. In the case of a surveillance system, consider that usage may peak during outbreaks, but remain moderate otherwise. Using orchestration software and modern cloud computing, resources can be acquired as necessary, minimizing computing costs. Workflows can similarly take advantage of container technology. Modern orchestration software is also tightly integrated with the DevOps model, a set of practices and tools to oversee development life cycle from implementation, all the way to deployment [77]. At the scale of services, this ensures that AlignDx technologies are consistently subject to testing, and validation practices, delivering quality code, leading to a better surveillance platform. The AlignDx workflow engine is more lenient in terms of this model in comparison to its service implementation, as any valid containerized pipeline can be used.

5.1.3 Avenues for Improvements

As a web-based platform, AlignDx is in a favorable position to take advantage of recent and coming technological advances. Consider the proliferation of smart devices, which has led to an increase in the volume and variety of health data, through accelerated data collection [78]. As these devices can communicate over a network, workflows can be designed to intake this data. This information can then be used to contextualize surveillance data. As these devices are not limited to human, but also animal and environmental health, these could potentially be utilized together in a one health surveillance model. This is an integrated approach to public health that seeks to combine human, animal and ecosystem health towards more-informed health solutions [15]. There are certainly existing systems that loosely follow this model, with varying degrees to implementation [15]. Although current AlignDx workflows are human-centric, the flexibility of the design makes it simple to design a novel workflow incorporating multi-domain surveillance data to generate comprehensive reports. While the system is currently focused on single workflows, there is also potential to chain multiple workflows together. Following the one health model with the input of multiple data streams, a workflow of interest could trigger contextual workflows. The commercialization of AI technologies, primarily through web APIs, could also be very useful for the platform. Consider the large language model Chat Generative Pre-trained Transformer (ChatGPT), which could play a pivotal role in human-computer interactions [79]. In the case of the surveillance workflow, this AI model could be used in combination with additional input parameters (sampling location, method, etc.) to construct comprehensive reports.

Although the workflow engine is an integral driver behind the AlignDx platform, it is a simple and novel approach with room for growth. From a system-wide perspective, pieces of the engine are fragmented, and could provide more functionality with tighter integration. These include workflow scheduling, execution, and monitoring, which are managed by the Celery, Factory and API services respectively. By joining these components end-to-end, the observability over the workflow process would increase, leading to greater user control over the entire surveillance analyses process. At the component level, changes in the current YAML structure for pipeline creation could lead to a more efficient system. This could range from including resource requirements for pipeline execution, typical runtimes, and more metadata for the related web-client form. Additionally, other file extensions could be explored for pipeline schemas, such as a native Python file format that provides greater integration in comparison to the YAML format. Alternatively, there are novel technologies within the workflow engine domain that could replace this component altogether.

5.2 Genomics Workflows

5.2.1 Using diverse data sources for better surveillance outcomes

Testing the metagenomic workflows with a wastewater and nasopharyngeal dataset provides critical information on their capabilities. From a surveillance perspective, these are two different methods of disease monitoring, each with their input source to output response. Together, they demonstrate the variety in valid data sources for the AlignDx genomic workflows, and the outcomes of their usage.

The surveillance workflow analysis of the wastewater dataset models an approach to environmental surveillance for disease monitoring [80]. Unlike the one-to-one nature of nasopharyngeal datasets, wastewater surveillance can dynamically monitor pathogen occurrence in communities [81]. This approach has been fruitful in early-detection of community-wide disease prevalence, such as with COVID-19 [82]. In the case of the wastewater dataset run in Chapter 3 by the Lookout workflow, SARS-related coronavirus was detected and reported at low abundance across two sampling sites. The presence of SARS-CoV-2 reads within these metagenomic samples is not unexpected, as it has previously been detected in patient fecal specimens, amongst other sites [83]. Additionally, detecting SARS-CoV-2 in wastewater treatment plant samples has been shown in numerous studies to correlate with diagnosed COVID-19 cases [84]. While SARS-CoV-2 was the only viral pathogen detected within this dataset, the diversity of profiled microbes suggests room for further analysis. Microbiome profiling for COVID-19 monitoring is an example where there is suggested co-occurrence of SARS-CoV-2 and certain microbiota [82]. Within this wastewater dataset run, *Pseudomonas spp.*, *Bacteroides spp.* and *Prevotella spp.* were some examples of observed positive correlating species with SARS-CoV-2 positive samples [82]. However, the low sequencing depth of this dataset makes it unsuitable for further analysis. This is amongst some of the many barriers currently impeding the use of wastewater surveillance in the public health response effectively. These include data uncertainty, measurement variability, method standardization, lack of expertise and resources, ethics and clear examples of utility [84]. AlignDx, in combination with the Lookout workflow, can address many of the post-sequencing issues, which similarly face these barriers. By curating this workflow further, or even developing a wastewater specific workflow, many of these barriers can be avoided by the end user. Thus, this provides a glimpse into the capabilities of wastewater as a data source for AlignDx genomic workflows.

While wastewater surveillance is more applicable to broad community surveillance, the nasopharyngeal approach allows for direct testing of individuals, and monitoring of their infections. This is in line with the standard approach to disease surveillance, which makes the clinical swab dataset suitable for performance evaluation of the genomic workflows. Nasopharyngeal swabs are a typical sampling site for respiratory pathogen diagnostics. Although optimal sampling sites may differ depending on the target microorganism, the nasopharynx is a principal colonization site for these pathogens [85]. With respiratory infections such as URIs and LRI on an upward trend, as detailed in Chapter 1, this dataset is good representative of modern surveillance interests. First, the genomic workflow was found to accurately predict SARS-CoV-2 in qPCR labelled samples (positive/negative). Accuracy is a common metric that generalizes the performance of a model by informing the reader on how well a prediction is correct but can be misleading. In the case of the clinical swab dataset, 16/66 samples are negatively labelled, demonstrating a major class imbalance, which in binary classification can greatly influence accuracy [86]. The impact of the classification threshold also greatly skews accuracy. Note that the qPCR analysis for SARS-CoV-2 genes used to generate “ground truth” labels on the dataset for performance evaluation does introduce some biases. Some studies have shown RT-qPCR sensitivity to vary greatly depending on component materials (buffer, reagents, etc.) [87]. Others have shown variation in diagnostic based on sampling site or even primer set [88]. These findings underline that all calculated performance metrics are only applicable for this dataset, under these conditions.

As these workflows run in a remote cloud environment, understanding the impact of sequencing depth on prediction accuracy within this dataset via subsampling could lead to performance benefits. This was done by calculating sensitivity and specificity, which have been shown in screening processes to be critical in performance assessment [89]. Reservoir subsampling was done in fixed and fractional values, where the former provides a consistent read size across samples, and the latter provides an accurate representation of the distribution. Correctly classifying the presence/absence of SARS-CoV-2 in the clinical swab dataset samples demonstrates a high degree of separability, even in deeply subsampled conditions. This suggests that the entire sample FASTQ file may not be necessary for correct classification of SARS-CoV-2 in this dataset, where 0.1% of the original sample size was sufficient for good performance. As internet upload speeds greatly vary depending on factors such as location, access and cost, decreasing the size of these datasets could greatly decrease analysis turnaround times.

5.2.2 Next Steps for AlignDx Genomics Workflows

The base pipeline used as the foundation of the genomic workflows presented in this thesis, alongside the target audience reports, make up the analysis and subsequent dissemination protocols for genomic AlignDx surveillance efforts. Using the Nextflow workflow system as the foundation for the various genomic workflows presented in Chapter 3 provides several advantages. As a mature workflow management system amongst its competitors, it provides readability, compactness, portability and provenance as core features [90]. These features are crucial in making workflow prototyping easy, which is integral in UCD. As this pipeline can function outside of the AlignDx system, adjustments can be made as required by the end user. From a workflow construction perspective, software can be developed, tested and pipelines can be validated independent of AlignDx with Nextflow. This modularity also means that the Lookout workflow can be open-sourced, and entirely community driven. The implementation of the base pipeline using Nextflow also demonstrates that any valid Nextflow/nf-core pipeline can be easily integrated into the platform. Consider that the nf-core community hosts a growing repository of 35 curated, open-source bioinformatic pipelines, each with a development team, publication, and release cycle, adhering to community guidelines [61]. Some of the available pipeline at the time of writing within the surveillance domain include variant calling, antimicrobial resistance, and dual host-pathogen analysis pipelines, amongst others [91]. While the scope of the Lookout workflow is pathogen identification via taxonomic profiling, these pipelines are some of the analyses that could replace or enhance it.

Using Kraken2 and Bracken within the Lookout workflow to taxonomically profile input data provides rapid classification, with good accuracy. Assessing the capabilities of classifiers is a complex process on its own, thus several studies have benchmarked these tools. Ye and colleagues benchmarked 20 classifiers, and found that Kraken2 performed consistently with the top classifiers and alongside Bracken, provided good accuracy in abundance profiling [92]. In viral pathogen detection, Kraken2 and other classifiers have shown comparable sensitivity and specificity to PCR based approaches [93]. This plays well with the Pathogen Panels DB, which focuses on virus taxonomy to simplify the complexities of taxonomic profiling in metagenomic datasets. One major impediment to pathogen detection from metagenomic data via Kraken 2 is that the reference k-mer database may not contain “truly” unique sequences. As explored by Doster et al., accurately identifying biologically relevant results is incumbent on the uniqueness of k-mers in the reference database [94]. This could be remedied by a larger database, although this will in turn increase resource requirements and turnaround times for

workflow execution. As this workflow is easily modified, alternative species profiling tools could be explored, such as marker gene approaches like mOTUs2 [95], or ML approaches like DeepMicrobes [96].

The metagenomic approach to pathogen surveillance provided by the AlignDx genomic workflows demonstrates the potential of sequencing for public health applications. Unlike traditional PCR based workflows, metagenomics can take an untargeted approach to detection. This is reflected in the surveillance workflow analysis of the wastewater dataset discussed above, where no targeted pathogens were suspected leading to surveillance with those samples. Metagenomics is also advantaged in its sensitivity, capability of detecting novel pathogens and co-infections, the depth of information it provides and the potential to shorten turnaround times [97]. While these make metagenomics an effective approach, there are certainly many limitations as well. In terms of viral metagenomics, as explored through the Lookout/surveillance workflow, detection sensitivity is a critical issue that can lead to false positives [97]. Furthermore, the high mutation rates common amongst viruses poses complications with taxonomic classification [97]. Addressing this becomes an issue of generating contextual data that aids in identifying biologically significant results [98]. In terms of implementation in the Lookout workflow, this could mean intaking this data as an input, or even generating it based on certain parameters. Additionally, steps such as quality filtering, or the removal of host reads may help, at the cost of computational resources and analysis speed [98]. Optimizing a bioinformatic pipeline for genomic based surveillance is a highly tailored endeavor, thus several studies have proposed methodology suited for these tasks. Buffet-Bataillon et al., proposed a quality-optimized process using Kraken and Bracken, improved by optimizing abundance thresholds, external controls and finally, a comprehensive classification reference database [99]. Overall, there are avenues for improvement with the AlignDx genomic workflows, improving surveillance prospects.

5.3 LFA Workflows

5.3.1 LFA test automation in AlignDx

Unlike the cost prohibitive nature and technical expertise required for genomic-based surveillance, the AlignDx VisuFlow workflow provides a POC LFA approach. LFA technologies are readily accessible by both healthcare workers and everyday individuals, at low cost with little expertise required for usage. By automating this process, VisuFlow demonstrates the utility of a POC approach to surveillance.

Effectively replicating the observation capabilities of the human eye through camera technology for LFA test automation requires testing with image quality variation. The VisuFlow training dataset presented in Chapter 4 tackles this using multiple LFA tests, under controlled lighting and dilution conditions. In a systematic review of COVID-19 LFA kits, test sensitivity was found to greatly vary depending on the manufacturer [100]. Capturing images from a variety of LFA kits is thus essential for modeling manufacturer differences. LFA technologies typically employ color, fluorescent or alternative labels, and the detection methods differ based on these labels [30]. Image processing of color based LFA tests kits, as used within the training dataset, is influenced by lighting conditions [101]. With varying light conditions, the optical signal retrieved by the capturing camera may differ significantly, which can lead to performance issues. This optical signal was also masked using serial dilution conditions, creating a greater diversity of images in the test dataset. Additionally, these dilutions, and ultimately capturing tests in ideal and sub-optimal conditions mock operator performance. Via the performance analysis, the VisuFlow workflow was found to accurately classify LFA tests based on manually generated labels. This was expected, as the classification discriminator of a line detection algorithm is relatively simple. As mentioned before, although accuracy helps generalize a model's performance, it is insufficient as a metric for binary classification, as done by VisuFlow. Multiple performance metrics were therefore calculated across temperature (lighting), dilution and block size parameters, where the model was found to be highly sensitive and specific in detection. Consistent performance across these three variables, and optimal and sub-optimal capturing conditions underlines its capabilities.

5.3.2 Next Steps for VisuFlow

While the training dataset does address variability in lighting conditions, the reproducibility of the LFA results using camera technology is impacted by numerous factors. These include differences in camera specifications, algorithm, and finally operating variations [102]. Standardizing capturing technology for LFA reading is difficult without a dedicated device and would likely increase manufacturing and buying costs. Park et al., proposed a general strategy to stabilize the target optical signal with a custom algorithm using a reference card with alignment marks, color standards and a QR code for spatial coordinate determination [102]. VisuFlow could incorporate a similar system by adjusting for a reference card, if available. However, the utility of a smartphone reader for a POC technology is in its convenience, therefore a software solution is preferable. In a study by Mendels et al., it was found that using convolutional neural network (CNN) model to classify 11 SARS-CoV-2 LFA tests provided 99.3% precision compared to eye [103]. Similarly, Turbé et al., demonstrated that CNN models had high sensitivity and specificity compared to interpretation by differently experienced health care workers [36]. Comprehensive artificial intelligence models, like CNNs may be suitable alternatives to the OpenCV ML algorithms used in the VisuFlow workflow.

A highlight of the VisuFlow workflow is its flexibility to both singular and mass image analysis. With singular image analysis, the benefit is to an everyday individual self-testing, providing an automated step to detection. As a standalone workflow, this is trivial, but when done for mass image analysis, with the archiving capabilities and mobile-first features of the AlignDx platform, this enhances the capabilities of LFA based surveillance. Consider the 2021 study by Lamb et al., which tested the performance of COVID-19 LFA kits on hospital workers, where results were submitted through an online portal, facilitating earlier detection of infection [104]. With some adjustments to allow data intake from multiple users, VisuFlow could be employed similarly. Furthermore, contact tracing could be examined via contextual metadata, which could aid in early response to disease spread. Analyte quantification is another facet of LFA surveillance that could be explored using the VisuFlow workflow, especially in mass surveillance scenarios. ML models have similarly been successful in this domain, where in color based LFA kits, quantification is calculated proportional to color intensity [105]. VisuFlow could also facilitate the analysis of multiplex LFA kits, such as those detecting multiple antigens. Alongside analyte quantification, these advanced LFA kits could be used to generate a comprehensive report summarizing the signal strength of each analyte.

References

1. Baker RE, Mahmud AS, Miller IF, Rajeev M, Rasambainarivo F, Rice BL, et al.. Infectious disease in an era of global change. *Nat Rev Microbiol*. Nature Publishing Group; 2022; doi: 10.1038/s41579-021-00639-z.
2. Spernovasilis N, Tsiodras S, Poulakou G. Emerging and Re-Emerging Infectious Diseases: Humankind's Companions and Competitors. *Microorganisms*. 2022; doi: 10.3390/microorganisms10010098.
3. Lajoie J. Understanding the measurement of global burden of disease. Winnipeg, Manitoba: National Collaborating Centre for Infectious Diseases;
4. Vos T, Lim SS, Abbafati C, Abbas KM, Abbasi M, Abbasifard M, et al.. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*. Elsevier; 2020; doi: 10.1016/S0140-6736(20)30925-9.
5. WHO: World health statistics 2022: monitoring health for the SDGs, sustainable development goals. <https://www.who.int/publications-detail-redirect/9789240051157> Accessed 2023 Mar 13.
6. Gibb R, Redding DW, Chin KQ, Donnelly CA, Blackburn TM, Newbold T, et al.. Zoonotic host diversity increases in human-dominated ecosystems. *Nature*. Nature Publishing Group; 2020; doi: 10.1038/s41586-020-2562-8.
7. Recht J, Schuenemann VJ, Sánchez-Villagra MR. Host Diversity and Origin of Zoonoses: The Ancient and the New. *Animals (Basel)*. 2020; doi: 10.3390/ani10091672.
8. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, et al.. Global trends in emerging infectious diseases. *Nature*. 2008; doi: 10.1038/nature06536.
9. Wang W-H, Thitithanyanont A, Urbina AN, Wang S-F. Emerging and Re-Emerging Diseases. *Pathogens*. 2021; doi: 10.3390/pathogens10070827.
10. Jin X, Ren J, Li R, Gao Y, Zhang H, Li J, et al.. Global burden of upper respiratory infections in 204 countries and territories, from 1990 to 2019. *eClinicalMedicine*. Elsevier; 2021; doi: 10.1016/j.eclinm.2021.100986.
11. Kyu HH, Vongpradith A, Sirota SB, Novotney A, Troeger CE, Doxey MC, et al.. Age–sex differences in the global burden of lower respiratory infections and risk factors, 1990–2019: results

- from the Global Burden of Disease Study 2019. *The Lancet Infectious Diseases*. 2022; doi: 10.1016/S1473-3099(22)00510-2.
12. Pires SM, Wyper GMA, Wengler A, Peñalvo JL, Haneef R, Moran D, et al.. Burden of Disease of COVID-19: Strengthening the Collaboration for National Studies. *Frontiers in Public Health*. 102022;
13. Mogharab V, Ostovar M, Ruszkowski J, Hussain SZM, Shrestha R, Yaqoob U, et al.. Global burden of the COVID-19 associated patient-related delay in emergency healthcare: a panel of systematic review and meta-analyses. *Globalization and Health*. 2022; doi: 10.1186/s12992-022-00836-2.
14. Threats I of M (US) F on M. Strategies for Disease Containment. Ethical and Legal Considerations in Mitigating Pandemic Disease: Workshop Summary. National Academies Press (US);
15. Bordier M, Uea-Anuwong T, Binot A, Hendriks P, Goutard FL. Characteristics of One Health surveillance systems: A systematic literature review. *Preventive Veterinary Medicine*. 2020; doi: 10.1016/j.prevetmed.2018.10.005.
16. Murray J, Cohen AL. Infectious Disease Surveillance. *International Encyclopedia of Public Health*. 2017; doi: 10.1016/B978-0-12-803678-5.00517-8.
17. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat Rev Genet*. Nature Publishing Group; 2018; doi: 10.1038/nrg.2017.88.
18. Nelson PP, Rath BA, Fragkou PC, Antalis E, Tsiodras S, Skevaki C. Current and Future Point-of-Care Tests for Emerging and New Respiratory Viruses and Future Perspectives. *Front Cell Infect Microbiol*. 2020; doi: 10.3389/fcimb.2020.00181.
19. Schmitz JE, Stratton CW, Persing DH, Tang Y-W. Forty Years of Molecular Diagnostics for Infectious Diseases. *Journal of Clinical Microbiology*. American Society for Microbiology; 2022; doi: 10.1128/jcm.02446-21.
20. Yang S, Rothman RE. PCR-based diagnostics for infectious diseases: uses, limitations, and future applications in acute-care settings. *Lancet Infect Dis*. 2004; doi: 10.1016/S1473-3099(04)01044-8.

21. Boudet A, Stephan R, Bravo S, Sasso M, Lavigne J-P. Limitation of Screening of Different Variants of SARS-CoV-2 by RT-PCR. *Diagnostics (Basel)*. 2021; doi: 10.3390/diagnostics11071241.
22. Vashist SK. Point-of-Care Diagnostics: Recent Advances and Trends. *Biosensors (Basel)*. 2017; doi: 10.3390/bios7040062.
23. Chiu CY, Miller SA. Clinical metagenomics. *Nat Rev Genet*. Nature Publishing Group; 2019; doi: 10.1038/s41576-019-0113-7.
24. Duan H, Li X, Mei A, Li P, Liu Y, Li X, et al.. The diagnostic value of metagenomic next-generation sequencing in infectious diseases. *BMC Infectious Diseases*. 2021; doi: 10.1186/s12879-020-05746-5.
25. Handel AS, Muller WJ, Planet PJ. Metagenomic Next-Generation Sequencing (mNGS): SARS-CoV-2 as an Example of the Technology's Potential Pediatric Infectious Disease Applications. *J Pediatric Infect Dis Soc*. 2021; doi: 10.1093/jpids/piab108.
26. Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview. *Human Immunology*. 2021; doi: 10.1016/j.humimm.2021.02.012.
27. Pearman WS, Freed NE, Silander OK. Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads. *BMC Bioinformatics*. 2020; doi: 10.1186/s12859-020-3528-4.
28. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*. 2014; doi: 10.1186/gb-2014-15-3-r46.
29. Fellows Yates JA, nf-core. (2021). MetroMap style pipeline workflow components.
30. Koczula KM, Gallotta A. Lateral flow assays. *Essays Biochem*. 2016; doi: 10.1042/EBC20150012.
31. Park H-D. Current Status of Clinical Application of Point-of-Care Testing. *Archives of Pathology & Laboratory Medicine*. 2020; doi: 10.5858/arpa.2020-0112-RA.
32. Zhang Y, Chai Y, Hu Z, Xu Z, Li M, Chen X, et al.. Recent Progress on Rapid Lateral Flow Assay-Based Early Diagnosis of COVID-19. *Frontiers in Bioengineering and Biotechnology*. 102022;

33. Peto T, Affron D, Afrough B, Agasu A, Ainsworth M, Allanson A, et al.. COVID-19: Rapid antigen detection for SARS-CoV-2 by lateral flow assay: A national systematic evaluation of sensitivity and specificity for mass-testing. *eClinicalMedicine*. Elsevier; 2021; doi: 10.1016/j.eclinm.2021.100924.
34. Dalrymple KA, Manner MD, Harmelink KA, Teska EP, Elison JT. An Examination of Recording Accuracy and Precision From Eye Tracking Data From Toddlerhood to Adulthood. *Frontiers in Psychology*. 2018;
35. Bosten JM. Do You See What I See? Diversity in Human Color Perception. *Annu Rev Vis Sci*. 2022; doi: 10.1146/annurev-vision-093020-112820.
36. Turbé V, Herbst C, Mngomezulu T, Meshkinfamfard S, Dlamini N, Mhlongo T, et al.. Deep learning of HIV field-based rapid tests. *Nat Med*. 2021; doi: 10.1038/s41591-021-01384-9.
37. Schary W, Paskali F, Rentschler S, Ruppert C, Wagner GE, Steinmetz I, et al.. Open-Source, Adaptable, All-in-One Smartphone-Based System for Quantitative Analysis of Point-of-Care Diagnostics. *Diagnostics (Basel)*. 2022; doi: 10.3390/diagnostics12030589.
38. Chappell K, Francou B, Habib C, Huby T, Leoni M, Cottin A, et al.. Galaxy Is a Suitable Bioinformatics Platform for the Molecular Diagnosis of Human Genetic Disorders Using High-Throughput Sequencing Data Analysis: Five Years of Experience in a Clinical Laboratory. *Clinical Chemistry*. 2022; doi: 10.1093/clinchem/hvab220.
39. Jalili V, Afgan E, Gu Q, Clements D, Blankenberg D, Goecks J, et al.. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Research*. 2020; doi: 10.1093/nar/gkaa434.
40. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al.. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. 2018; doi: 10.1093/bioinformatics/bty407.
41. Pais RJ. Predictive Modelling in Clinical Bioinformatics: Key Concepts for Startups. *BioTech (Basel)*. 2022; doi: 10.3390/biotech11030035.
42. Pavelin K, Cham JA, de Matos P, Brooksbank C, Cameron G, Steinbeck C. Bioinformatics Meets User-Centred Design: A Perspective. *PLoS Comput Biol*. 2012; doi: 10.1371/journal.pcbi.1002554.

43. S. A, Iñiguez-Jarrín C, Lopez O, Gonzalez-Ibea D, Pérez-Román E, Borredà C, et al.. Applying User Centred Design to Improve the Design of Genomic User Interfaces: *Proceedings of the 16th International Conference on Evaluation of Novel Approaches to Software Engineering*. SCITEPRESS - Science and Technology Publications;
44. Wratten L, Wilm A, Göke J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nat Methods*. Nature Publishing Group; 2021; doi: 10.1038/s41592-021-01254-9.
45. Sanderson L-A, Caron CT, Tan RL, Bett KE. A PostgreSQL Tripal solution for large-scale genotypic and phenotypic data. *Database (Oxford)*. 2021; doi: 10.1093/database/baab051.
46. Merkel D. Docker: lightweight Linux containers for consistent development and deployment. *Linux Journal*. 2014;
47. Hykes S. (2023). Docker Compose (Version 2.18.1). <https://github.com/docker/compose>.
48. Walke J. (2022). React.js (Version 18.2.0). <https://github.com/facebook/react>.
49. Vercel. (2022). Next.js (Version 13.0.1). <https://github.com/vercel/next.js>.
50. Crockford D. (2017). JSON (Version ECMA-404). <https://www.json.org/json-en.html>.
51. Ramírez S. (2022). FastAPI (Version 0.85.0). <https://github.com/tiangolo/fastapi>.
52. Colvin S. (2022). Pydantic (Version 1.10.4). <https://github.com/pydantic/pydantic>.
53. IBM: What is a REST API? <https://www.ibm.com/topics/rest-apis> Accessed 2023 Mar 31.
54. MDN: HTTP request methods. <https://developer.mozilla.org/en-US/docs/Web/HTTP/Methods> (2023). Accessed 2023 Mar 31.
55. Tus. (2023). tUSD (Version 1.11). <https://github.com/tus/tUSD>.
56. Stonebreaker M. (2019). PostgreSQL (Version 12). <https://github.com/postgres/postgres>
57. Sanfilippo S. (2022). Redis. (Version 7). <https://github.com/redis/redis>.
58. Solem A. (2022). Celery (Version 5.2.7). <https://github.com/celery/celery>.
59. Net I döt, Evans C, Ben-Kiki O. (2021). YAML Ain't Markup Language (Version 1.2.2). <https://yaml.org/spec/1.2.2/>.

60. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. *Nat Biotechnol.* Nature Publishing Group; 2017; doi: 10.1038/nbt.3820.
61. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al.. The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol.* Nature Publishing Group; 2020; doi: 10.1038/s41587-020-0439-x.
62. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* 2010; doi: 10.1093/nar/gkp1137.
63. Li H. (2018). lh3/seqtk: Toolkit for processing sequences in FASTA/Q formats (Version 1.3). <https://github.com/lh3/seqtk>.
64. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biology.* 2019; doi: 10.1186/s13059-019-1891-0.
65. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput Sci.* PeerJ Inc.; 2017; doi: 10.7717/peerj-cs.104.
66. Parmer J, Parmer C, Sundquist M, Johnson A. (2023). Plotly (Version 5.13.0). <https://github.com/plotly/plotly.py>
67. Ewels P, Magnusson M, Lundin S, Källner M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics.* 2016; doi: 10.1093/bioinformatics/btw354.
68. McKinney W. (2022). Pandas (Version 1.4.1). <https://pandas.pydata.org/>
69. Karr AF, Hauzel J, Porter AA, Schaefer M. Measuring quality of DNA sequence data via degradation. *PLOS ONE.* Public Library of Science; 2022; doi: 10.1371/journal.pone.0271970.
70. Dórea FC, Revie CW. Data-Driven Surveillance: Effective Collection, Integration, and Interpretation of Data to Support Decision Making. *Frontiers in Veterinary Science.* 82021;
71. Mustafa U, Kreppel KS, Brinkel J, Sauli E. Digital Technologies to Enhance Infectious Disease Surveillance in Tanzania: A Scoping Review. *Healthcare.* Multidisciplinary Digital Publishing Institute; 2023; doi: 10.3390/healthcare11040470.

72. McGraw D, Mandl KD. Privacy protections to encourage use of health-relevant digital data in a learning health system. *npj Digit Med*. Nature Publishing Group; 2021; doi: 10.1038/s41746-020-00362-8.
73. Bentotahewa V, Hewage C, Williams J. Solutions to Big Data Privacy and Security Challenges Associated With COVID-19 Surveillance Systems. *Frontiers in Big Data*. 42021;
74. Ling-Hu T, Rios-Guzman E, Lorenzo-Redondo R, Ozer EA, Hultquist JF. Challenges and Opportunities for Global Genomic Surveillance Strategies in the COVID-19 Era. *Viruses*. 2022; doi: 10.3390/v14112532.
75. Lage R, Dolog P, Leginus M. The Role of Adaptive Elements in Web-Based Surveillance System User Interfaces. In: Dimitrova V, Kuflik T, Chin D, Ricci F, Dolog P, Houben G-J, editors. *User Modeling, Adaptation, and Personalization*. Cham: Springer International Publishing;
76. Singh A, Triulzi G, Magee CL. Technological improvement rate predictions for all technologies: Use of patent data and an extended domain description. *Research Policy*. 2021; doi: 10.1016/j.respol.2021.104294.
77. Kadri S, Sboner A, Sigaras A, Roy S. Containers in Bioinformatics: Applications, Practical Considerations, and Best Practices in Molecular Pathology. *The Journal of Molecular Diagnostics*. 2022; doi: 10.1016/j.jmoldx.2022.01.006.
78. Sahu KS, Majowicz SE, Dubin JA, Morita PP. NextGen Public Health Surveillance and the Internet of Things (IoT). *Frontiers in Public Health*. 92021;
79. Xue VW, Lei P, Cho WC. The potential impact of ChatGPT in clinical and translational medicine. *Clin Transl Med*. 2023; doi: 10.1002/ctm2.1216.
80. Kilaru P, Hill D, Anderson K, Collins MB, Green H, Kmush BL, et al.. Wastewater Surveillance for Infectious Disease: A Systematic Review. *American Journal of Epidemiology*. 2023; doi: 10.1093/aje/kwac175.
81. Sinclair RG, Choi CY, Riley MR, Gerba CP. Pathogen Surveillance Through Monitoring of Sewer Systems. *Adv Appl Microbiol*. 2008; doi: 10.1016/S0065-2164(08)00609-6.
82. Brumfield KD, Leddy M, Usmani M, Cotruvo JA, Tien C-T, Dorsey S, et al.. Microbiome Analysis for Wastewater Surveillance during COVID-19. *mBio*. doi: 10.1128/mbio.00591-22.

83. Wang W, Xu Y, Gao R, Lu R, Han K, Wu G, et al.. Detection of SARS-CoV-2 in Different Types of Clinical Specimens. *JAMA*. 2020; doi: 10.1001/jama.2020.3786.
84. McClary-Gutierrez JS, Mattioli MC, Marcenac P, Silverman AI, Boehm AB, Bibby K, et al.. SARS-CoV-2 Wastewater Surveillance for Public Health Action. *Emerg Infect Dis*. 2021; doi: 10.3201/eid2709.210753.
85. Edouard S, Million M, Bachar D, Dubourg G, Michelle C, Ninove L, et al.. The nasopharyngeal microbiota in patients with viral respiratory tract infections is enriched in bacterial pathogens. *Eur J Clin Microbiol Infect Dis*. 2018; doi: 10.1007/s10096-018-3305-8.
86. Luque A, Carrasco A, Martín A, de las Heras A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*. 2019; doi: 10.1016/j.patcog.2019.02.023.
87. Alcoba-Florez J, Gil-Campesino H, Artola DG-M de, González-Montelongo R, Valenzuela-Fernández A, Ciuffreda L, et al.. Sensitivity of different RT-qPCR solutions for SARS-CoV-2 detection. *Int J Infect Dis*. 2020; doi: 10.1016/j.ijid.2020.07.058.
88. Lawrence Panchali MJ, Oh HJ, Lee YM, Kim C-M, Tariq M, Seo J-W, et al.. Accuracy of Real-Time Polymerase Chain Reaction in COVID-19 Patients. *Microbiol Spectr*. doi: 10.1128/spectrum.00591-21.
89. Trevelyan R. Sensitivity, Specificity, and Predictive Values: Foundations, Plausibilities, and Pitfalls in Research and Practice. *Front Public Health*. 2017; doi: 10.3389/fpubh.2017.00307.
90. Ahmed AE, Allen JM, Bhat T, Burra P, Fliege CE, Hart SN, et al.. Design considerations for workflow management systems use in production genomics research and the clinic. *Sci Rep*. Nature Publishing Group; 2021; doi: 10.1038/s41598-021-99288-8.
91. Nf-Core: Nf-Core Pipelines. <https://nf-co.re/pipelines> Accessed 2023 Apr 24.
92. Ye SH, Siddle KJ, Park DJ, Sabeti PC. Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell*. 2019; doi: 10.1016/j.cell.2019.07.010.
93. Carbo EC, Sidorov IA, van Rijn-Klink AL, Pappas N, van Boheemen S, Mei H, et al.. Performance of Five Metagenomic Classifiers for Virus Pathogen Detection Using Respiratory Samples from a Clinical Cohort. *Pathogens*. 2022; doi: 10.3390/pathogens11030340.

94. Doster E, Rovira P, Noyes NR, Burgess BA, Yang X, Weinroth MD, et al.. A Cautionary Report for Pathogen Identification Using Shotgun Metagenomics; A Comparison to Aerobic Culture and Polymerase Chain Reaction for *Salmonella enterica* Identification. *Front Microbiol.* 2019; doi: 10.3389/fmicb.2019.02499.
95. Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh H-J, Cuenca M, et al.. Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat Commun.* Nature Publishing Group; 2019; doi: 10.1038/s41467-019-08844-4.
96. Liang Q, Bible PW, Liu Y, Zou B, Wei L. DeepMicrobes: taxonomic classification for metagenomics with deep learning. *NAR Genomics and Bioinformatics.* 2020; doi: 10.1093/nargab/lqaa009.
97. Nooij S, Schmitz D, Vennema H, Kroneman A, Koopmans MPG. Overview of Virus Metagenomic Classification Methods and Their Biological Applications. *Frontiers in Microbiology.* 92018;
98. Maljkovic Berry I, Melendrez MC, Bishop-Lilly KA, Rutvisuttinunt W, Pollett S, Talundzic E, et al.. Next Generation Sequencing and Bioinformatics Methodologies for Infectious Disease Research and Public Health: Approaches, Applications, and Considerations for Development of Laboratory Capacity. *The Journal of Infectious Diseases.* 2020; doi: 10.1093/infdis/jiz286.
99. Marais G, Hardie D, Brink A. A case for investment in clinical metagenomics in low-income and middle-income countries. *The Lancet Microbe.* 2023; doi: 10.1016/S2666-5247(22)00328-7.
100. Mistry DA, Wang JY, Moeser M-E, Starkey T, Lee LYW. A systematic review of the sensitivity and specificity of lateral flow devices in the detection of SARS-CoV-2. *BMC Infect Dis.* 2021; doi: 10.1186/s12879-021-06528-3.
101. Sajid M, Kawde A-N, Daud M. Designs, formats and applications of lateral flow assay: A literature review. *Journal of Saudi Chemical Society.* 2015; doi: 10.1016/j.jscs.2014.09.001.
102. Park J-H, Park E-K, Cho YK, Shin I-S, Lee H. Normalizing the Optical Signal Enables Robust Assays with Lateral Flow Biosensors. *ACS Omega.* 2022; doi: 10.1021/acsomega.2c00793.
103. Mendels D-A, Dortet L, Emerald C, Oueslati S, Girlich D, Ronat J-B, et al.. Using artificial intelligence to improve COVID-19 rapid diagnostic test result interpretation. *Proc Natl Acad Sci U S A.* 2021; doi: 10.1073/pnas.2019893118.

104. Lamb G, Heskin J, Randell P, Mughal N, Moore LS, Jones R, et al.. Real-world evaluation of COVID-19 lateral flow device (LFD) mass-testing in healthcare workers at a London hospital; a prospective cohort analysis. *J Infect.* 2021; doi: 10.1016/j.jinf.2021.07.038.
105. Foyosal KH, Seo SE, Kim MJ, Kwon OS, Chong JW. Analyte Quantity Detection from Lateral Flow Assay Using a Smartphone. *Sensors (Basel).* 2019; doi: 10.3390/s19214812.

Appendix A – Supplementary Data

Metric Calculations:

Hyp = Hyperparameter

TP = True Positive

FP = False Positive

TN = True Negative

FN = False Negative

$$\text{Accuracy (ACC)} = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

$$\text{True Positive Rate (TPR) OR Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{False Positive Rate (FPR) OR Fall-Out} = \frac{FP}{FP + TN}$$

$$\text{True Negative Rate (TNR) OR Specificity} = \frac{TN}{TN + FP}$$

$$\text{Positive Predictive Value (PPV) OR Precision} = \frac{TP}{TP + FP}$$

$$\text{False Negative Rate (FNR)} = \frac{FN}{TP + FN}$$

$$\text{Negative Predictive Value (NPV)} = \frac{TN}{TN + FN}$$

$$\text{False Discovery Rate (FDR)} = \frac{FP}{TP + FP}$$

$$F1 = 2 * \frac{\text{Precision} * \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$



Figure A.1 - Cluster map of the top 100 represented species post-human-filtering. Samples are ordered as follows: 1-16 (Negative), 17-32 (Outpatient), 33-48 (non-ICU), 49-66 (ICU).

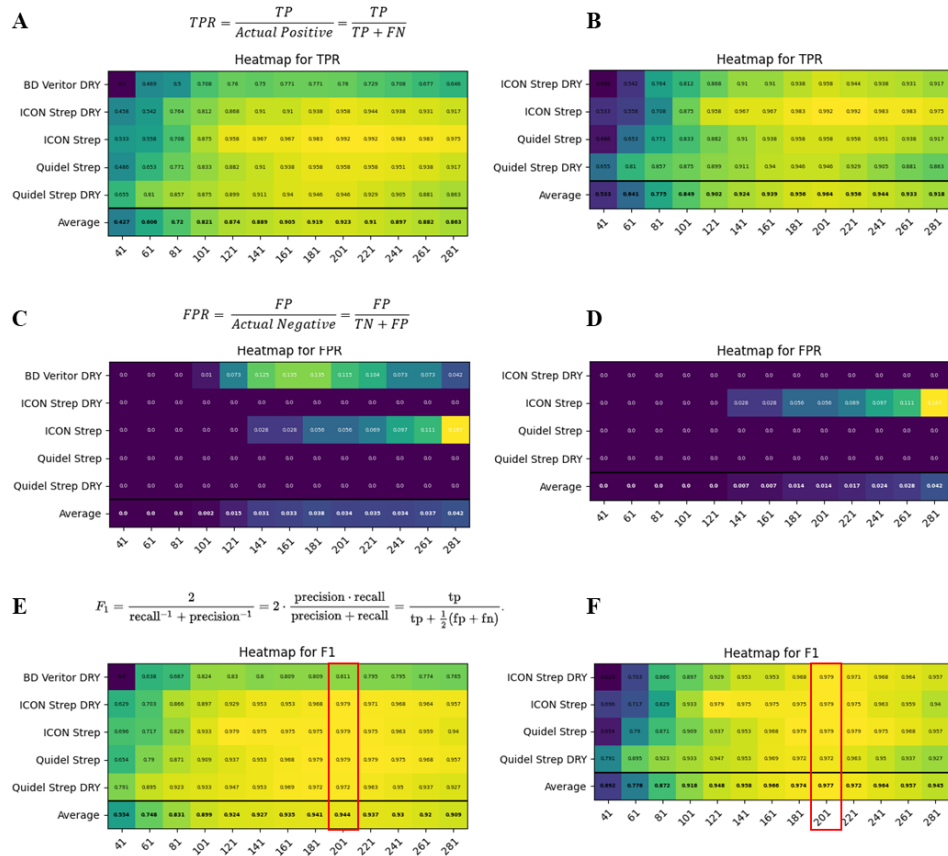


Figure A.2 - Determination of True Positive Rate (TPR), False Positive Rate (FPR), and F1 statistic under variable block sizes with and without BD Veritor. TPR (A), FPR (C) and F1 (E) statistics were calculated for all five lateral flow assays under a range of block sizes to reveal the relative effectiveness of the algorithm on detecting true positives and true negatives. The BD Veritor test performed poorly for all statistics and was removed from a second round of analyses performed (B, D, and F). For both analyses (with or without the BD Veritor data included), the block size of 201 (red boxes – E and F) produced the greatest F1 statistic value and was used in subsequent analyses.

Table A.1 - SARS-CoV-2 100% sampling performance metrics. Thresholds were generated based on Bracken relative abundance measurements. ACC: Accuracy, TPR: True Positive Rate, TNR: True Negative Rate, PPV: Positive Predictive Value, NPV: Negative Predictive Value, FPR: False Positive Rate, FNR: False Negative Rate, FDR: False Discovery Rate, TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative.

Threshold	ACC	TPR	TNR	PPV	NPV	FPR	FNR	FDR	TP	FP	FN	TN
0	0.76	1.00	0.00	0.76	0.00	1.00	0.00	0.24	50	16	0	0
0.00001	0.82	1.00	0.25	0.81	1.00	0.75	0.00	0.19	50	12	0	4
0.00002	0.88	1.00	0.50	0.86	1.00	0.50	0.00	0.14	50	8	0	8
0.00003	0.89	1.00	0.56	0.88	1.00	0.44	0.00	0.12	50	7	0	9
0.00004	0.91	1.00	0.63	0.89	1.00	0.38	0.00	0.11	50	6	0	10
0.00005	0.92	1.00	0.69	0.91	1.00	0.31	0.00	0.09	50	5	0	11
0.00007	0.94	1.00	0.75	0.93	1.00	0.25	0.00	0.07	50	4	0	12
0.00011	0.95	1.00	0.81	0.94	1.00	0.19	0.00	0.06	50	3	0	13
0.00022	0.94	0.98	0.81	0.94	0.93	0.19	0.02	0.06	49	3	1	13
0.00041	0.95	0.98	0.88	0.96	0.93	0.13	0.02	0.04	49	2	1	14
0.00044	0.97	0.98	0.94	0.98	0.94	0.06	0.02	0.02	49	1	1	15
0.00047	0.95	0.96	0.94	0.98	0.88	0.06	0.04	0.02	48	1	2	15
0.00056	0.97	0.96	1.00	1.00	0.89	0.00	0.04	0.00	48	0	2	16
0.00097	0.95	0.94	1.00	1.00	0.84	0.00	0.06	0.00	47	0	3	16
0.00124	0.94	0.92	1.00	1.00	0.80	0.00	0.08	0.00	46	0	4	16
0.00325	0.92	0.90	1.00	1.00	0.76	0.00	0.10	0.00	45	0	5	16
0.00333	0.91	0.88	1.00	1.00	0.73	0.00	0.12	0.00	44	0	6	16
0.00358	0.89	0.86	1.00	1.00	0.70	0.00	0.14	0.00	43	0	7	16
0.00369	0.88	0.84	1.00	1.00	0.67	0.00	0.16	0.00	42	0	8	16
0.00393	0.86	0.82	1.00	1.00	0.64	0.00	0.18	0.00	41	0	9	16
0.00745	0.85	0.80	1.00	1.00	0.62	0.00	0.20	0.00	40	0	10	16
0.00763	0.83	0.78	1.00	1.00	0.59	0.00	0.22	0.00	39	0	11	16
0.01515	0.82	0.76	1.00	1.00	0.57	0.00	0.24	0.00	38	0	12	16
0.01996	0.80	0.74	1.00	1.00	0.55	0.00	0.26	0.00	37	0	13	16
0.03026	0.79	0.72	1.00	1.00	0.53	0.00	0.28	0.00	36	0	14	16
0.03069	0.77	0.70	1.00	1.00	0.52	0.00	0.30	0.00	35	0	15	16
0.05662	0.76	0.68	1.00	1.00	0.50	0.00	0.32	0.00	34	0	16	16
0.06142	0.74	0.66	1.00	1.00	0.48	0.00	0.34	0.00	33	0	17	16
0.06469	0.73	0.64	1.00	1.00	0.47	0.00	0.36	0.00	32	0	18	16
0.07608	0.71	0.62	1.00	1.00	0.46	0.00	0.38	0.00	31	0	19	16
0.0889	0.70	0.60	1.00	1.00	0.44	0.00	0.40	0.00	30	0	20	16
0.11984	0.68	0.58	1.00	1.00	0.43	0.00	0.42	0.00	29	0	21	16
0.12344	0.67	0.56	1.00	1.00	0.42	0.00	0.44	0.00	28	0	22	16
0.12946	0.65	0.54	1.00	1.00	0.41	0.00	0.46	0.00	27	0	23	16
0.12984	0.64	0.52	1.00	1.00	0.40	0.00	0.48	0.00	26	0	24	16
0.14051	0.62	0.50	1.00	1.00	0.39	0.00	0.50	0.00	25	0	25	16

0.15734	0.61	0.48	1.00	1.00	0.38	0.00	0.52	0.00	24	0	26	16
0.17765	0.59	0.46	1.00	1.00	0.37	0.00	0.54	0.00	23	0	27	16
0.36115	0.58	0.44	1.00	1.00	0.36	0.00	0.56	0.00	22	0	28	16
0.36996	0.56	0.42	1.00	1.00	0.36	0.00	0.58	0.00	21	0	29	16
0.37412	0.55	0.40	1.00	1.00	0.35	0.00	0.60	0.00	20	0	30	16
0.38499	0.53	0.38	1.00	1.00	0.34	0.00	0.62	0.00	19	0	31	16
0.38679	0.52	0.36	1.00	1.00	0.33	0.00	0.64	0.00	18	0	32	16
0.40892	0.50	0.34	1.00	1.00	0.33	0.00	0.66	0.00	17	0	33	16
0.4298	0.48	0.32	1.00	1.00	0.32	0.00	0.68	0.00	16	0	34	16
0.45204	0.47	0.30	1.00	1.00	0.31	0.00	0.70	0.00	15	0	35	16
0.46285	0.45	0.28	1.00	1.00	0.31	0.00	0.72	0.00	14	0	36	16
0.51992	0.44	0.26	1.00	1.00	0.30	0.00	0.74	0.00	13	0	37	16
0.54689	0.42	0.24	1.00	1.00	0.30	0.00	0.76	0.00	12	0	38	16
0.59298	0.41	0.22	1.00	1.00	0.29	0.00	0.78	0.00	11	0	39	16
0.61561	0.39	0.20	1.00	1.00	0.29	0.00	0.80	0.00	10	0	40	16
0.63223	0.38	0.18	1.00	1.00	0.28	0.00	0.82	0.00	9	0	41	16
0.66594	0.36	0.16	1.00	1.00	0.28	0.00	0.84	0.00	8	0	42	16
0.67502	0.35	0.14	1.00	1.00	0.27	0.00	0.86	0.00	7	0	43	16
0.68086	0.33	0.12	1.00	1.00	0.27	0.00	0.88	0.00	6	0	44	16
0.71742	0.32	0.10	1.00	1.00	0.26	0.00	0.90	0.00	5	0	45	16
0.79497	0.30	0.08	1.00	1.00	0.26	0.00	0.92	0.00	4	0	46	16
0.86445	0.29	0.06	1.00	1.00	0.25	0.00	0.94	0.00	3	0	47	16
0.93554	0.27	0.04	1.00	1.00	0.25	0.00	0.96	0.00	2	0	48	16
0.93734	0.26	0.02	1.00	1.00	0.25	0.00	0.98	0.00	1	0	49	16

Table A.2 - SARS-CoV-2 10% sampling performance metrics. Thresholds were generated based on Bracken relative abundance measurements. ACC: Accuracy, TPR: True Positive Rate, TNR: True Negative Rate, PPV: Positive Predictive Value, NPV: Negative Predictive Value, FPR: False Positive Rate, FNR: False Negative Rate, FDR: False Discovery Rate, TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative.

Threshold	ACC	TPR	TNR	PPV	NPV	FPR	FNR	FDR	TP	FP	FN	TN
0	0.76	1.00	0.00	0.76	0.00	1.00	0.00	0.24	50	16	0	0
2.00E-05	0.92	1.00	0.69	0.91	1.00	0.31	0.00	0.09	50	5	0	11
4.00E-05	0.94	1.00	0.75	0.93	1.00	0.25	0.00	0.07	50	4	0	12
0.0001	0.95	1.00	0.81	0.94	1.00	0.19	0.00	0.06	50	3	0	13
0.00023	0.94	0.98	0.81	0.94	0.93	0.19	0.02	0.06	49	3	1	13
0.00043	0.95	0.98	0.88	0.96	0.93	0.13	0.02	0.04	49	2	1	14
0.00045	0.94	0.96	0.88	0.96	0.88	0.13	0.04	0.04	48	2	2	14
0.00055	0.95	0.96	0.94	0.98	0.88	0.06	0.04	0.02	48	1	2	15
0.00096	0.95	0.94	1.00	1.00	0.84	0.00	0.06	0.00	47	0	3	16
0.00121	0.94	0.92	1.00	1.00	0.80	0.00	0.08	0.00	46	0	4	16
0.00328	0.92	0.90	1.00	1.00	0.76	0.00	0.10	0.00	45	0	5	16
0.00363	0.91	0.88	1.00	1.00	0.73	0.00	0.12	0.00	44	0	6	16
0.00374	0.89	0.86	1.00	1.00	0.70	0.00	0.14	0.00	43	0	7	16
0.00375	0.88	0.84	1.00	1.00	0.67	0.00	0.16	0.00	42	0	8	16
0.0038	0.86	0.82	1.00	1.00	0.64	0.00	0.18	0.00	41	0	9	16
0.00752	0.85	0.80	1.00	1.00	0.62	0.00	0.20	0.00	40	0	10	16
0.00814	0.83	0.78	1.00	1.00	0.59	0.00	0.22	0.00	39	0	11	16
0.01546	0.82	0.76	1.00	1.00	0.57	0.00	0.24	0.00	38	0	12	16
0.02154	0.80	0.74	1.00	1.00	0.55	0.00	0.26	0.00	37	0	13	16
0.03042	0.79	0.72	1.00	1.00	0.53	0.00	0.28	0.00	36	0	14	16
0.03149	0.77	0.70	1.00	1.00	0.52	0.00	0.30	0.00	35	0	15	16
0.05744	0.76	0.68	1.00	1.00	0.50	0.00	0.32	0.00	34	0	16	16
0.06256	0.74	0.66	1.00	1.00	0.48	0.00	0.34	0.00	33	0	17	16
0.06436	0.73	0.64	1.00	1.00	0.47	0.00	0.36	0.00	32	0	18	16
0.07618	0.71	0.62	1.00	1.00	0.46	0.00	0.38	0.00	31	0	19	16
0.08956	0.70	0.60	1.00	1.00	0.44	0.00	0.40	0.00	30	0	20	16
0.12473	0.68	0.58	1.00	1.00	0.43	0.00	0.42	0.00	29	0	21	16
0.12684	0.67	0.56	1.00	1.00	0.42	0.00	0.44	0.00	28	0	22	16
0.13198	0.65	0.54	1.00	1.00	0.41	0.00	0.46	0.00	27	0	23	16
0.13387	0.64	0.52	1.00	1.00	0.40	0.00	0.48	0.00	26	0	24	16
0.14339	0.62	0.50	1.00	1.00	0.39	0.00	0.50	0.00	25	0	25	16
0.15866	0.61	0.48	1.00	1.00	0.38	0.00	0.52	0.00	24	0	26	16
0.17903	0.59	0.46	1.00	1.00	0.37	0.00	0.54	0.00	23	0	27	16
0.36953	0.58	0.44	1.00	1.00	0.36	0.00	0.56	0.00	22	0	28	16
0.37292	0.56	0.42	1.00	1.00	0.36	0.00	0.58	0.00	21	0	29	16
0.38116	0.55	0.40	1.00	1.00	0.35	0.00	0.60	0.00	20	0	30	16

0.39211	0.53	0.38	1.00	1.00	0.34	0.00	0.62	0.00	19	0	31	16
0.40649	0.52	0.36	1.00	1.00	0.33	0.00	0.64	0.00	18	0	32	16
0.41328	0.50	0.34	1.00	1.00	0.33	0.00	0.66	0.00	17	0	33	16
0.43336	0.48	0.32	1.00	1.00	0.32	0.00	0.68	0.00	16	0	34	16
0.46657	0.47	0.30	1.00	1.00	0.31	0.00	0.70	0.00	15	0	35	16
0.47476	0.45	0.28	1.00	1.00	0.31	0.00	0.72	0.00	14	0	36	16
0.52362	0.44	0.26	1.00	1.00	0.30	0.00	0.74	0.00	13	0	37	16
0.56686	0.42	0.24	1.00	1.00	0.30	0.00	0.76	0.00	12	0	38	16
0.59722	0.41	0.22	1.00	1.00	0.29	0.00	0.78	0.00	11	0	39	16
0.62015	0.39	0.20	1.00	1.00	0.29	0.00	0.80	0.00	10	0	40	16
0.64221	0.38	0.18	1.00	1.00	0.28	0.00	0.82	0.00	9	0	41	16
0.67611	0.36	0.16	1.00	1.00	0.28	0.00	0.84	0.00	8	0	42	16
0.6845	0.35	0.14	1.00	1.00	0.27	0.00	0.86	0.00	7	0	43	16
0.68585	0.33	0.12	1.00	1.00	0.27	0.00	0.88	0.00	6	0	44	16
0.72171	0.32	0.10	1.00	1.00	0.26	0.00	0.90	0.00	5	0	45	16
0.79715	0.30	0.08	1.00	1.00	0.26	0.00	0.92	0.00	4	0	46	16
0.86747	0.29	0.06	1.00	1.00	0.25	0.00	0.94	0.00	3	0	47	16
0.93715	0.27	0.04	1.00	1.00	0.25	0.00	0.96	0.00	2	0	48	16
0.94141	0.26	0.02	1.00	1.00	0.25	0.00	0.98	0.00	1	0	49	16

Table A.3 - SARS-CoV-2 1% sampling performance metrics. Thresholds were generated based on Bracken relative abundance measurements. ACC: Accuracy, TPR: True Positive Rate, TNR: True Negative Rate, PPV: Positive Predictive Value, NPV: Negative Predictive Value, FPR: False Positive Rate, FNR: False Negative Rate, FDR: False Discovery Rate, TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative.

Threshold	ACC	TPR	TNR	PPV	NPV	FPR	FNR	FDR	TP	FP	FN	TN
0	0.76	1.00	0.00	0.76	0.00	1.00	0.00	0.24	50	16	0	0
0.00056	0.95	0.94	1.00	1.00	0.84	0.00	0.06	0.00	47	0	3	16
0.00117	0.94	0.92	1.00	1.00	0.80	0.00	0.08	0.00	46	0	4	16
0.00323	0.92	0.90	1.00	1.00	0.76	0.00	0.10	0.00	45	0	5	16
0.00336	0.91	0.88	1.00	1.00	0.73	0.00	0.12	0.00	44	0	6	16
0.00357	0.89	0.86	1.00	1.00	0.70	0.00	0.14	0.00	43	0	7	16
0.00378	0.88	0.84	1.00	1.00	0.67	0.00	0.16	0.00	42	0	8	16
0.00482	0.86	0.82	1.00	1.00	0.64	0.00	0.18	0.00	41	0	9	16
0.00787	0.85	0.80	1.00	1.00	0.62	0.00	0.20	0.00	40	0	10	16
0.0085	0.83	0.78	1.00	1.00	0.59	0.00	0.22	0.00	39	0	11	16
0.01707	0.82	0.76	1.00	1.00	0.57	0.00	0.24	0.00	38	0	12	16
0.02398	0.80	0.74	1.00	1.00	0.55	0.00	0.26	0.00	37	0	13	16
0.03098	0.79	0.72	1.00	1.00	0.53	0.00	0.28	0.00	36	0	14	16
0.03374	0.77	0.70	1.00	1.00	0.52	0.00	0.30	0.00	35	0	15	16
0.06126	0.76	0.68	1.00	1.00	0.50	0.00	0.32	0.00	34	0	16	16
0.06508	0.74	0.66	1.00	1.00	0.48	0.00	0.34	0.00	33	0	17	16
0.06795	0.73	0.64	1.00	1.00	0.47	0.00	0.36	0.00	32	0	18	16
0.07625	0.71	0.62	1.00	1.00	0.46	0.00	0.38	0.00	31	0	19	16
0.09257	0.70	0.60	1.00	1.00	0.44	0.00	0.40	0.00	30	0	20	16
0.12807	0.68	0.58	1.00	1.00	0.43	0.00	0.42	0.00	29	0	21	16
0.12902	0.67	0.56	1.00	1.00	0.42	0.00	0.44	0.00	28	0	22	16
0.13328	0.65	0.54	1.00	1.00	0.41	0.00	0.46	0.00	27	0	23	16
0.14542	0.64	0.52	1.00	1.00	0.40	0.00	0.48	0.00	26	0	24	16
0.14613	0.62	0.50	1.00	1.00	0.39	0.00	0.50	0.00	25	0	25	16
0.15775	0.61	0.48	1.00	1.00	0.38	0.00	0.52	0.00	24	0	26	16
0.18284	0.59	0.46	1.00	1.00	0.37	0.00	0.54	0.00	23	0	27	16
0.37812	0.58	0.44	1.00	1.00	0.36	0.00	0.56	0.00	22	0	28	16
0.38008	0.56	0.42	1.00	1.00	0.36	0.00	0.58	0.00	21	0	29	16
0.38431	0.55	0.40	1.00	1.00	0.35	0.00	0.60	0.00	20	0	30	16
0.39936	0.53	0.38	1.00	1.00	0.34	0.00	0.62	0.00	19	0	31	16
0.42168	0.52	0.36	1.00	1.00	0.33	0.00	0.64	0.00	18	0	32	16
0.43995	0.50	0.34	1.00	1.00	0.33	0.00	0.66	0.00	17	0	33	16
0.44249	0.48	0.32	1.00	1.00	0.32	0.00	0.68	0.00	16	0	34	16
0.48022	0.47	0.30	1.00	1.00	0.31	0.00	0.70	0.00	15	0	35	16
0.52567	0.45	0.28	1.00	1.00	0.31	0.00	0.72	0.00	14	0	36	16
0.53122	0.44	0.26	1.00	1.00	0.30	0.00	0.74	0.00	13	0	37	16

0.60019	0.42	0.24	1.00	1.00	0.30	0.00	0.76	0.00	12	0	38	16
0.6027	0.41	0.22	1.00	1.00	0.29	0.00	0.78	0.00	11	0	39	16
0.62558	0.39	0.20	1.00	1.00	0.29	0.00	0.80	0.00	10	0	40	16
0.65347	0.38	0.18	1.00	1.00	0.28	0.00	0.82	0.00	9	0	41	16
0.69107	0.36	0.16	1.00	1.00	0.28	0.00	0.84	0.00	8	0	42	16
0.69334	0.35	0.14	1.00	1.00	0.27	0.00	0.86	0.00	7	0	43	16
0.69844	0.33	0.12	1.00	1.00	0.27	0.00	0.88	0.00	6	0	44	16
0.72537	0.32	0.10	1.00	1.00	0.26	0.00	0.90	0.00	5	0	45	16
0.80108	0.30	0.08	1.00	1.00	0.26	0.00	0.92	0.00	4	0	46	16
0.86992	0.29	0.06	1.00	1.00	0.25	0.00	0.94	0.00	3	0	47	16
0.94195	0.27	0.04	1.00	1.00	0.25	0.00	0.96	0.00	2	0	48	16
0.94935	0.26	0.02	1.00	1.00	0.25	0.00	0.98	0.00	1	0	49	16

Table A.4 - SARS-CoV-2 0.1% sampling performance metrics. Thresholds were generated based on Bracken relative abundance measurements. ACC: Accuracy, TPR: True Positive Rate, TNR: True Negative Rate, PPV: Positive Predictive Value, NPV: Negative Predictive Value, FPR: False Positive Rate, FNR: False Negative Rate, FDR: False Discovery Rate, TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative.

Threshold	ACC	TPR	TNR	PPV	NPV	FPR	FNR	FDR	TP	FP	FN	TN
0	0.76	1.00	0.00	0.76	0.00	1.00	0.00	0.24	50	16	0	0
0.0008	0.91	0.88	1.00	1.00	0.73	0.00	0.12	0.00	44	0	6	16
0.00098	0.89	0.86	1.00	1.00	0.70	0.00	0.14	0.00	43	0	7	16
0.00207	0.88	0.84	1.00	1.00	0.67	0.00	0.16	0.00	42	0	8	16
0.00467	0.86	0.82	1.00	1.00	0.64	0.00	0.18	0.00	41	0	9	16
0.00934	0.85	0.80	1.00	1.00	0.62	0.00	0.20	0.00	40	0	10	16
0.01082	0.83	0.78	1.00	1.00	0.59	0.00	0.22	0.00	39	0	11	16
0.01299	0.82	0.76	1.00	1.00	0.57	0.00	0.24	0.00	38	0	12	16
0.02526	0.80	0.74	1.00	1.00	0.55	0.00	0.26	0.00	37	0	13	16
0.03088	0.79	0.72	1.00	1.00	0.53	0.00	0.28	0.00	36	0	14	16
0.03412	0.77	0.70	1.00	1.00	0.52	0.00	0.30	0.00	35	0	15	16
0.0651	0.76	0.68	1.00	1.00	0.50	0.00	0.32	0.00	34	0	16	16
0.06533	0.74	0.66	1.00	1.00	0.48	0.00	0.34	0.00	33	0	17	16
0.07274	0.73	0.64	1.00	1.00	0.47	0.00	0.36	0.00	32	0	18	16
0.08441	0.71	0.62	1.00	1.00	0.46	0.00	0.38	0.00	31	0	19	16
0.09054	0.70	0.60	1.00	1.00	0.44	0.00	0.40	0.00	30	0	20	16
0.1256	0.68	0.58	1.00	1.00	0.43	0.00	0.42	0.00	29	0	21	16
0.13316	0.67	0.56	1.00	1.00	0.42	0.00	0.44	0.00	28	0	22	16
0.14758	0.65	0.54	1.00	1.00	0.41	0.00	0.46	0.00	27	0	23	16
0.16086	0.64	0.52	1.00	1.00	0.40	0.00	0.48	0.00	26	0	24	16
0.17061	0.62	0.50	1.00	1.00	0.39	0.00	0.50	0.00	25	0	25	16
0.1847	0.61	0.48	1.00	1.00	0.38	0.00	0.52	0.00	24	0	26	16
0.18482	0.59	0.46	1.00	1.00	0.37	0.00	0.54	0.00	23	0	27	16
0.39372	0.58	0.44	1.00	1.00	0.36	0.00	0.56	0.00	22	0	28	16
0.40655	0.56	0.42	1.00	1.00	0.36	0.00	0.58	0.00	21	0	29	16
0.40987	0.55	0.40	1.00	1.00	0.35	0.00	0.60	0.00	20	0	30	16
0.41106	0.53	0.38	1.00	1.00	0.34	0.00	0.62	0.00	19	0	31	16
0.43021	0.52	0.36	1.00	1.00	0.33	0.00	0.64	0.00	18	0	32	16
0.47674	0.50	0.34	1.00	1.00	0.33	0.00	0.66	0.00	17	0	33	16
0.49465	0.48	0.32	1.00	1.00	0.32	0.00	0.68	0.00	16	0	34	16
0.5367	0.47	0.30	1.00	1.00	0.31	0.00	0.70	0.00	15	0	35	16
0.55014	0.45	0.28	1.00	1.00	0.31	0.00	0.72	0.00	14	0	36	16
0.61286	0.44	0.26	1.00	1.00	0.30	0.00	0.74	0.00	13	0	37	16
0.64019	0.42	0.24	1.00	1.00	0.30	0.00	0.76	0.00	12	0	38	16
0.64493	0.41	0.22	1.00	1.00	0.29	0.00	0.78	0.00	11	0	39	16
0.67221	0.39	0.20	1.00	1.00	0.29	0.00	0.80	0.00	10	0	40	16

0.71417	0.38	0.18	1.00	1.00	0.28	0.00	0.82	0.00	9	0	41	16
0.71768	0.36	0.16	1.00	1.00	0.28	0.00	0.84	0.00	8	0	42	16
0.71885	0.35	0.14	1.00	1.00	0.27	0.00	0.86	0.00	7	0	43	16
0.72868	0.33	0.12	1.00	1.00	0.27	0.00	0.88	0.00	6	0	44	16
0.77497	0.32	0.10	1.00	1.00	0.26	0.00	0.90	0.00	5	0	45	16
0.8073	0.30	0.08	1.00	1.00	0.26	0.00	0.92	0.00	4	0	46	16
0.87598	0.29	0.06	1.00	1.00	0.25	0.00	0.94	0.00	3	0	47	16
0.95751	0.27	0.04	1.00	1.00	0.25	0.00	0.96	0.00	2	0	48	16
0.96454	0.26	0.02	1.00	1.00	0.25	0.00	0.98	0.00	1	0	49	16

Table A.5 - SARS-CoV-2 0.01% sampling performance metrics. Thresholds were generated based on Bracken relative abundance measurements. ACC: Accuracy, TPR: True Positive Rate, TNR: True Negative Rate, PPV: Positive Predictive Value, NPV: Negative Predictive Value, FPR: False Positive Rate, FNR: False Negative Rate, FDR: False Discovery Rate, TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative.

Threshold	ACC	TPR	TNR	PPV	NPV	FPR	FNR	FDR	TP	FP	FN	TN
0	0.76	1.00	0.00	0.76	0.00	1.00	0.00	0.24	50	16	0	0
0.06955	0.74	0.66	1.00	1.00	0.48	0.00	0.34	0.00	33	0	17	16
0.07018	0.73	0.64	1.00	1.00	0.47	0.00	0.36	0.00	32	0	18	16
0.07336	0.71	0.62	1.00	1.00	0.46	0.00	0.38	0.00	31	0	19	16
0.08556	0.70	0.60	1.00	1.00	0.44	0.00	0.40	0.00	30	0	20	16
0.11852	0.68	0.58	1.00	1.00	0.43	0.00	0.42	0.00	29	0	21	16
0.18243	0.67	0.56	1.00	1.00	0.42	0.00	0.44	0.00	28	0	22	16
0.2	0.65	0.54	1.00	1.00	0.41	0.00	0.46	0.00	27	0	23	16
0.20548	0.64	0.52	1.00	1.00	0.40	0.00	0.48	0.00	26	0	24	16
0.22656	0.62	0.50	1.00	1.00	0.39	0.00	0.50	0.00	25	0	25	16
0.2381	0.61	0.48	1.00	1.00	0.38	0.00	0.52	0.00	24	0	26	16
0.3871	0.59	0.46	1.00	1.00	0.37	0.00	0.54	0.00	23	0	27	16
0.41573	0.58	0.44	1.00	1.00	0.36	0.00	0.56	0.00	22	0	28	16
0.46541	0.56	0.42	1.00	1.00	0.36	0.00	0.58	0.00	21	0	29	16
0.48903	0.55	0.40	1.00	1.00	0.35	0.00	0.60	0.00	20	0	30	16
0.52607	0.53	0.38	1.00	1.00	0.34	0.00	0.62	0.00	19	0	31	16
0.56329	0.52	0.36	1.00	1.00	0.33	0.00	0.64	0.00	18	0	32	16
0.6378	0.50	0.34	1.00	1.00	0.33	0.00	0.66	0.00	17	0	33	16
0.65805	0.48	0.32	1.00	1.00	0.32	0.00	0.68	0.00	16	0	34	16
0.67516	0.47	0.30	1.00	1.00	0.31	0.00	0.70	0.00	15	0	35	16
0.68834	0.45	0.28	1.00	1.00	0.31	0.00	0.72	0.00	14	0	36	16
0.74479	0.44	0.26	1.00	1.00	0.30	0.00	0.74	0.00	13	0	37	16
0.76098	0.42	0.24	1.00	1.00	0.30	0.00	0.76	0.00	12	0	38	16
0.76943	0.41	0.22	1.00	1.00	0.29	0.00	0.78	0.00	11	0	39	16
0.82154	0.39	0.20	1.00	1.00	0.29	0.00	0.80	0.00	10	0	40	16
0.83217	0.38	0.18	1.00	1.00	0.28	0.00	0.82	0.00	9	0	41	16
0.84422	0.36	0.16	1.00	1.00	0.28	0.00	0.84	0.00	8	0	42	16
0.85903	0.35	0.14	1.00	1.00	0.27	0.00	0.86	0.00	7	0	43	16
0.93402	0.33	0.12	1.00	1.00	0.27	0.00	0.88	0.00	6	0	44	16
1	0.32	0.10	1.00	1.00	0.26	0.00	0.90	0.00	5	0	45	16

Table A.6 - SARS-CoV-2 10 million read sampling performance metrics. Thresholds were generated based on Bracken relative abundance measurements. ACC: Accuracy, TPR: True Positive Rate, TNR: True Negative Rate, PPV: Positive Predictive Value, NPV: Negative Predictive Value, FPR: False Positive Rate, FNR: False Negative Rate, FDR: False Discovery Rate, TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative.

Threshold	ACC	TPR	TNR	PPV	NPV	FPR	FNR	FDR	TP	FP	FN	TN
0	0.76	1.00	0.00	0.76	0.00	1.00	0.00	0.24	50	16	0	0
2.00E-05	0.92	1.00	0.69	0.91	1.00	0.31	0.00	0.09	50	5	0	11
9.00E-05	0.94	1.00	0.75	0.93	1.00	0.25	0.00	0.07	50	4	0	12
0.00012	0.95	1.00	0.81	0.94	1.00	0.19	0.00	0.06	50	3	0	13
0.00019	0.94	0.98	0.81	0.94	0.93	0.19	0.02	0.06	49	3	1	13
0.00044	0.95	0.98	0.88	0.96	0.93	0.13	0.02	0.04	49	2	1	14
0.00045	0.97	0.98	0.94	0.98	0.94	0.06	0.02	0.02	49	1	1	15
0.00047	0.95	0.96	0.94	0.98	0.88	0.06	0.04	0.02	48	1	2	15
0.00055	0.97	0.96	1.00	1.00	0.89	0.00	0.04	0.00	48	0	2	16
0.00079	0.95	0.94	1.00	1.00	0.84	0.00	0.06	0.00	47	0	3	16
0.00124	0.94	0.92	1.00	1.00	0.80	0.00	0.08	0.00	46	0	4	16
0.00335	0.92	0.90	1.00	1.00	0.76	0.00	0.10	0.00	45	0	5	16
0.00338	0.91	0.88	1.00	1.00	0.73	0.00	0.12	0.00	44	0	6	16
0.00358	0.89	0.86	1.00	1.00	0.70	0.00	0.14	0.00	43	0	7	16
0.00381	0.88	0.84	1.00	1.00	0.67	0.00	0.16	0.00	42	0	8	16
0.00415	0.86	0.82	1.00	1.00	0.64	0.00	0.18	0.00	41	0	9	16
0.00735	0.85	0.80	1.00	1.00	0.62	0.00	0.20	0.00	40	0	10	16
0.00767	0.83	0.78	1.00	1.00	0.59	0.00	0.22	0.00	39	0	11	16
0.01565	0.82	0.76	1.00	1.00	0.57	0.00	0.24	0.00	38	0	12	16
0.02082	0.80	0.74	1.00	1.00	0.55	0.00	0.26	0.00	37	0	13	16
0.03029	0.79	0.72	1.00	1.00	0.53	0.00	0.28	0.00	36	0	14	16
0.03148	0.77	0.70	1.00	1.00	0.52	0.00	0.30	0.00	35	0	15	16
0.05779	0.76	0.68	1.00	1.00	0.50	0.00	0.32	0.00	34	0	16	16
0.06136	0.74	0.66	1.00	1.00	0.48	0.00	0.34	0.00	33	0	17	16
0.06462	0.73	0.64	1.00	1.00	0.47	0.00	0.36	0.00	32	0	18	16
0.07629	0.71	0.62	1.00	1.00	0.46	0.00	0.38	0.00	31	0	19	16
0.08976	0.70	0.60	1.00	1.00	0.44	0.00	0.40	0.00	30	0	20	16
0.12318	0.68	0.58	1.00	1.00	0.43	0.00	0.42	0.00	29	0	21	16
0.12687	0.67	0.56	1.00	1.00	0.42	0.00	0.44	0.00	28	0	22	16
0.13118	0.65	0.54	1.00	1.00	0.41	0.00	0.46	0.00	27	0	23	16
0.13411	0.64	0.52	1.00	1.00	0.40	0.00	0.48	0.00	26	0	24	16
0.14239	0.62	0.50	1.00	1.00	0.39	0.00	0.50	0.00	25	0	25	16
0.15815	0.61	0.48	1.00	1.00	0.38	0.00	0.52	0.00	24	0	26	16
0.17817	0.59	0.46	1.00	1.00	0.37	0.00	0.54	0.00	23	0	27	16
0.36823	0.58	0.44	1.00	1.00	0.36	0.00	0.56	0.00	22	0	28	16
0.37197	0.56	0.42	1.00	1.00	0.36	0.00	0.58	0.00	21	0	29	16

0.37853	0.55	0.40	1.00	1.00	0.35	0.00	0.60	0.00	20	0	30	16
0.39184	0.53	0.38	1.00	1.00	0.34	0.00	0.62	0.00	19	0	31	16
0.39801	0.52	0.36	1.00	1.00	0.33	0.00	0.64	0.00	18	0	32	16
0.41439	0.50	0.34	1.00	1.00	0.33	0.00	0.66	0.00	17	0	33	16
0.43217	0.48	0.32	1.00	1.00	0.32	0.00	0.68	0.00	16	0	34	16
0.46444	0.47	0.30	1.00	1.00	0.31	0.00	0.70	0.00	15	0	35	16
0.46792	0.45	0.28	1.00	1.00	0.31	0.00	0.72	0.00	14	0	36	16
0.52266	0.44	0.26	1.00	1.00	0.30	0.00	0.74	0.00	13	0	37	16
0.56261	0.42	0.24	1.00	1.00	0.30	0.00	0.76	0.00	12	0	38	16
0.59642	0.41	0.22	1.00	1.00	0.29	0.00	0.78	0.00	11	0	39	16
0.6205	0.39	0.20	1.00	1.00	0.29	0.00	0.80	0.00	10	0	40	16
0.64139	0.38	0.18	1.00	1.00	0.28	0.00	0.82	0.00	9	0	41	16
0.6722	0.36	0.16	1.00	1.00	0.28	0.00	0.84	0.00	8	0	42	16
0.68258	0.35	0.14	1.00	1.00	0.27	0.00	0.86	0.00	7	0	43	16
0.68566	0.33	0.12	1.00	1.00	0.27	0.00	0.88	0.00	6	0	44	16
0.72237	0.32	0.10	1.00	1.00	0.26	0.00	0.90	0.00	5	0	45	16
0.79635	0.30	0.08	1.00	1.00	0.26	0.00	0.92	0.00	4	0	46	16
0.86612	0.29	0.06	1.00	1.00	0.25	0.00	0.94	0.00	3	0	47	16
0.93728	0.27	0.04	1.00	1.00	0.25	0.00	0.96	0.00	2	0	48	16
0.9402	0.26	0.02	1.00	1.00	0.25	0.00	0.98	0.00	1	0	49	16

Table A.7 - SARS-CoV-2 1 million read sampling performance metrics. Thresholds were generated based on Bracken relative abundance measurements. ACC: Accuracy, TPR: True Positive Rate, TNR: True Negative Rate, PPV: Positive Predictive Value, NPV: Negative Predictive Value, FPR: False Positive Rate, FNR: False Negative Rate, FDR: False Discovery Rate, TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative.

Threshold	ACC	TPR	TNR	PPV	NPV	FPR	FNR	FDR	TP	FP	FN	TN
0	0.76	1.00	0.00	0.76	0.00	1.00	0.00	0.24	50	16	0	0
0.00049	0.98	0.98	1.00	1.00	0.94	0.00	0.02	0.00	49	0	1	16
0.00052	0.97	0.96	1.00	1.00	0.89	0.00	0.04	0.00	48	0	2	16
0.00112	0.95	0.94	1.00	1.00	0.84	0.00	0.06	0.00	47	0	3	16
0.00128	0.94	0.92	1.00	1.00	0.80	0.00	0.08	0.00	46	0	4	16
0.00312	0.92	0.90	1.00	1.00	0.76	0.00	0.10	0.00	45	0	5	16
0.00347	0.91	0.88	1.00	1.00	0.73	0.00	0.12	0.00	44	0	6	16
0.00371	0.89	0.86	1.00	1.00	0.70	0.00	0.14	0.00	43	0	7	16
0.00375	0.88	0.84	1.00	1.00	0.67	0.00	0.16	0.00	42	0	8	16
0.00408	0.86	0.82	1.00	1.00	0.64	0.00	0.18	0.00	41	0	9	16
0.00785	0.85	0.80	1.00	1.00	0.62	0.00	0.20	0.00	40	0	10	16
0.00807	0.83	0.78	1.00	1.00	0.59	0.00	0.22	0.00	39	0	11	16
0.01506	0.82	0.76	1.00	1.00	0.57	0.00	0.24	0.00	38	0	12	16
0.0222	0.80	0.74	1.00	1.00	0.55	0.00	0.26	0.00	37	0	13	16
0.0311	0.79	0.72	1.00	1.00	0.53	0.00	0.28	0.00	36	0	14	16
0.03321	0.77	0.70	1.00	1.00	0.52	0.00	0.30	0.00	35	0	15	16
0.05961	0.76	0.68	1.00	1.00	0.50	0.00	0.32	0.00	34	0	16	16
0.06392	0.74	0.66	1.00	1.00	0.48	0.00	0.34	0.00	33	0	17	16
0.06598	0.73	0.64	1.00	1.00	0.47	0.00	0.36	0.00	32	0	18	16
0.0751	0.71	0.62	1.00	1.00	0.46	0.00	0.38	0.00	31	0	19	16
0.09212	0.70	0.60	1.00	1.00	0.44	0.00	0.40	0.00	30	0	20	16
0.1253	0.68	0.58	1.00	1.00	0.43	0.00	0.42	0.00	29	0	21	16
0.12856	0.67	0.56	1.00	1.00	0.42	0.00	0.44	0.00	28	0	22	16
0.13398	0.65	0.54	1.00	1.00	0.41	0.00	0.46	0.00	27	0	23	16
0.13637	0.64	0.52	1.00	1.00	0.40	0.00	0.48	0.00	26	0	24	16
0.14443	0.62	0.50	1.00	1.00	0.39	0.00	0.50	0.00	25	0	25	16
0.15992	0.61	0.48	1.00	1.00	0.38	0.00	0.52	0.00	24	0	26	16
0.18227	0.59	0.46	1.00	1.00	0.37	0.00	0.54	0.00	23	0	27	16
0.37528	0.58	0.44	1.00	1.00	0.36	0.00	0.56	0.00	22	0	28	16
0.37654	0.56	0.42	1.00	1.00	0.36	0.00	0.58	0.00	21	0	29	16
0.38703	0.55	0.40	1.00	1.00	0.35	0.00	0.60	0.00	20	0	30	16
0.40051	0.53	0.38	1.00	1.00	0.34	0.00	0.62	0.00	19	0	31	16
0.41911	0.52	0.36	1.00	1.00	0.33	0.00	0.64	0.00	18	0	32	16
0.42179	0.50	0.34	1.00	1.00	0.33	0.00	0.66	0.00	17	0	33	16
0.44053	0.48	0.32	1.00	1.00	0.32	0.00	0.68	0.00	16	0	34	16
0.4735	0.47	0.30	1.00	1.00	0.31	0.00	0.70	0.00	15	0	35	16

0.52389	0.45	0.28	1.00	1.00	0.31	0.00	0.72	0.00	14	0	36	16
0.52525	0.44	0.26	1.00	1.00	0.30	0.00	0.74	0.00	13	0	37	16
0.5972	0.42	0.24	1.00	1.00	0.30	0.00	0.76	0.00	12	0	38	16
0.60366	0.41	0.22	1.00	1.00	0.29	0.00	0.78	0.00	11	0	39	16
0.62605	0.39	0.20	1.00	1.00	0.29	0.00	0.80	0.00	10	0	40	16
0.65565	0.38	0.18	1.00	1.00	0.28	0.00	0.82	0.00	9	0	41	16
0.68563	0.36	0.16	1.00	1.00	0.28	0.00	0.84	0.00	8	0	42	16
0.69195	0.35	0.14	1.00	1.00	0.27	0.00	0.86	0.00	7	0	43	16
0.69487	0.33	0.12	1.00	1.00	0.27	0.00	0.88	0.00	6	0	44	16
0.72677	0.32	0.10	1.00	1.00	0.26	0.00	0.90	0.00	5	0	45	16
0.80043	0.30	0.08	1.00	1.00	0.26	0.00	0.92	0.00	4	0	46	16
0.86926	0.29	0.06	1.00	1.00	0.25	0.00	0.94	0.00	3	0	47	16
0.9396	0.27	0.04	1.00	1.00	0.25	0.00	0.96	0.00	2	0	48	16
0.94601	0.26	0.02	1.00	1.00	0.25	0.00	0.98	0.00	1	0	49	16

Table A.8 - SARS-CoV-2 100 thousand read sampling performance metrics. Thresholds were generated based on Bracken relative abundance measurements. ACC: Accuracy, TPR: True Positive Rate, TNR: True Negative Rate, PPV: Positive Predictive Value, NPV: Negative Predictive Value, FPR: False Positive Rate, FNR: False Negative Rate, FDR: False Discovery Rate, TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative.

Threshold	ACC	TPR	TNR	PPV	NPV	FPR	FNR	FDR	TP	FP	FN	TN
0	0.76	1.00	0.00	0.76	0.00	1.00	0.00	0.24	50	16	0	0
0.0006	0.91	0.88	1.00	1.00	0.73	0.00	0.12	0.00	44	0	6	16
0.00107	0.89	0.86	1.00	1.00	0.70	0.00	0.14	0.00	43	0	7	16
0.00256	0.88	0.84	1.00	1.00	0.67	0.00	0.16	0.00	42	0	8	16
0.0044	0.86	0.82	1.00	1.00	0.64	0.00	0.18	0.00	41	0	9	16
0.00697	0.85	0.80	1.00	1.00	0.62	0.00	0.20	0.00	40	0	10	16
0.00866	0.83	0.78	1.00	1.00	0.59	0.00	0.22	0.00	39	0	11	16
0.01398	0.82	0.76	1.00	1.00	0.57	0.00	0.24	0.00	38	0	12	16
0.02802	0.80	0.74	1.00	1.00	0.55	0.00	0.26	0.00	37	0	13	16
0.02907	0.79	0.72	1.00	1.00	0.53	0.00	0.28	0.00	36	0	14	16
0.03082	0.77	0.70	1.00	1.00	0.52	0.00	0.30	0.00	35	0	15	16
0.06097	0.76	0.68	1.00	1.00	0.50	0.00	0.32	0.00	34	0	16	16
0.06491	0.74	0.66	1.00	1.00	0.48	0.00	0.34	0.00	33	0	17	16
0.07233	0.73	0.64	1.00	1.00	0.47	0.00	0.36	0.00	32	0	18	16
0.07446	0.71	0.62	1.00	1.00	0.46	0.00	0.38	0.00	31	0	19	16
0.09905	0.70	0.60	1.00	1.00	0.44	0.00	0.40	0.00	30	0	20	16
0.13878	0.68	0.58	1.00	1.00	0.43	0.00	0.42	0.00	29	0	21	16
0.1403	0.67	0.56	1.00	1.00	0.42	0.00	0.44	0.00	28	0	22	16
0.14133	0.65	0.54	1.00	1.00	0.41	0.00	0.46	0.00	27	0	23	16
0.14924	0.64	0.52	1.00	1.00	0.40	0.00	0.48	0.00	26	0	24	16
0.16017	0.62	0.50	1.00	1.00	0.39	0.00	0.50	0.00	25	0	25	16
0.16823	0.61	0.48	1.00	1.00	0.38	0.00	0.52	0.00	24	0	26	16
0.18477	0.59	0.46	1.00	1.00	0.37	0.00	0.54	0.00	23	0	27	16
0.38531	0.58	0.44	1.00	1.00	0.36	0.00	0.56	0.00	22	0	28	16
0.38902	0.56	0.42	1.00	1.00	0.36	0.00	0.58	0.00	21	0	29	16
0.39598	0.55	0.40	1.00	1.00	0.35	0.00	0.60	0.00	20	0	30	16
0.41573	0.53	0.38	1.00	1.00	0.34	0.00	0.62	0.00	19	0	31	16
0.42431	0.52	0.36	1.00	1.00	0.33	0.00	0.64	0.00	18	0	32	16
0.45641	0.50	0.34	1.00	1.00	0.33	0.00	0.66	0.00	17	0	33	16
0.48716	0.48	0.32	1.00	1.00	0.32	0.00	0.68	0.00	16	0	34	16
0.5086	0.47	0.30	1.00	1.00	0.31	0.00	0.70	0.00	15	0	35	16
0.53505	0.45	0.28	1.00	1.00	0.31	0.00	0.72	0.00	14	0	36	16
0.58333	0.44	0.26	1.00	1.00	0.30	0.00	0.74	0.00	13	0	37	16
0.60939	0.42	0.24	1.00	1.00	0.30	0.00	0.76	0.00	12	0	38	16
0.63861	0.41	0.22	1.00	1.00	0.29	0.00	0.78	0.00	11	0	39	16
0.67296	0.39	0.20	1.00	1.00	0.29	0.00	0.80	0.00	10	0	40	16

0.70936	0.38	0.18	1.00	1.00	0.28	0.00	0.82	0.00	9	0	41	16
0.71229	0.36	0.16	1.00	1.00	0.28	0.00	0.84	0.00	8	0	42	16
0.71583	0.35	0.14	1.00	1.00	0.27	0.00	0.86	0.00	7	0	43	16
0.72024	0.33	0.12	1.00	1.00	0.27	0.00	0.88	0.00	6	0	44	16
0.74534	0.32	0.10	1.00	1.00	0.26	0.00	0.90	0.00	5	0	45	16
0.80741	0.30	0.08	1.00	1.00	0.26	0.00	0.92	0.00	4	0	46	16
0.87093	0.29	0.06	1.00	1.00	0.25	0.00	0.94	0.00	3	0	47	16
0.94724	0.27	0.04	1.00	1.00	0.25	0.00	0.96	0.00	2	0	48	16
0.9536	0.26	0.02	1.00	1.00	0.25	0.00	0.98	0.00	1	0	49	16

Table A.9 - SARS-CoV-2 10 thousand read sampling performance metrics. Thresholds were generated based on Bracken relative abundance measurements. ACC: Accuracy, TPR: True Positive Rate, TNR: True Negative Rate, PPV: Positive Predictive Value, NPV: Negative Predictive Value, FPR: False Positive Rate, FNR: False Negative Rate, FDR: False Discovery Rate, TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative.

Threshold	ACC	TPR	TNR	PPV	NPV	FPR	FNR	FDR	TP	FP	FN	TN
0	0.76	1.00	0.00	0.76	0.00	1.00	0.00	0.24	50	16	0	0
0.02364	0.79	0.72	1.00	1.00	0.53	0.00	0.28	0.00	36	0	14	16
0.03129	0.77	0.70	1.00	1.00	0.52	0.00	0.30	0.00	35	0	15	16
0.05666	0.76	0.68	1.00	1.00	0.50	0.00	0.32	0.00	34	0	16	16
0.06891	0.74	0.66	1.00	1.00	0.48	0.00	0.34	0.00	33	0	17	16
0.09524	0.73	0.64	1.00	1.00	0.47	0.00	0.36	0.00	32	0	18	16
0.1405	0.71	0.62	1.00	1.00	0.46	0.00	0.38	0.00	31	0	19	16
0.1435	0.70	0.60	1.00	1.00	0.44	0.00	0.40	0.00	30	0	20	16
0.18437	0.68	0.58	1.00	1.00	0.43	0.00	0.42	0.00	29	0	21	16
0.18519	0.67	0.56	1.00	1.00	0.42	0.00	0.44	0.00	28	0	22	16
0.188	0.65	0.54	1.00	1.00	0.41	0.00	0.46	0.00	27	0	23	16
0.2013	0.64	0.52	1.00	1.00	0.40	0.00	0.48	0.00	26	0	24	16
0.22807	0.62	0.50	1.00	1.00	0.39	0.00	0.50	0.00	25	0	25	16
0.23256	0.61	0.48	1.00	1.00	0.38	0.00	0.52	0.00	24	0	26	16
0.37037	0.59	0.46	1.00	1.00	0.37	0.00	0.54	0.00	23	0	27	16
0.46724	0.58	0.44	1.00	1.00	0.36	0.00	0.56	0.00	22	0	28	16
0.47706	0.56	0.42	1.00	1.00	0.36	0.00	0.58	0.00	21	0	29	16
0.48889	0.55	0.40	1.00	1.00	0.35	0.00	0.60	0.00	20	0	30	16
0.5	0.53	0.38	1.00	1.00	0.34	0.00	0.62	0.00	19	0	31	16
0.54634	0.52	0.36	1.00	1.00	0.33	0.00	0.64	0.00	18	0	32	16
0.56632	0.50	0.34	1.00	1.00	0.33	0.00	0.66	0.00	17	0	33	16
0.59364	0.48	0.32	1.00	1.00	0.32	0.00	0.68	0.00	16	0	34	16
0.64535	0.47	0.30	1.00	1.00	0.31	0.00	0.70	0.00	15	0	35	16
0.69223	0.45	0.28	1.00	1.00	0.31	0.00	0.72	0.00	14	0	36	16
0.70455	0.44	0.26	1.00	1.00	0.30	0.00	0.74	0.00	13	0	37	16
0.75	0.42	0.24	1.00	1.00	0.30	0.00	0.76	0.00	12	0	38	16
0.75556	0.41	0.22	1.00	1.00	0.29	0.00	0.78	0.00	11	0	39	16
0.75728	0.39	0.20	1.00	1.00	0.29	0.00	0.80	0.00	10	0	40	16
0.77477	0.38	0.18	1.00	1.00	0.28	0.00	0.82	0.00	9	0	41	16
0.82097	0.36	0.16	1.00	1.00	0.28	0.00	0.84	0.00	8	0	42	16
0.82168	0.35	0.14	1.00	1.00	0.27	0.00	0.86	0.00	7	0	43	16
0.85934	0.33	0.12	1.00	1.00	0.27	0.00	0.88	0.00	6	0	44	16
0.92902	0.32	0.10	1.00	1.00	0.26	0.00	0.90	0.00	5	0	45	16
0.9665	0.30	0.08	1.00	1.00	0.26	0.00	0.92	0.00	4	0	46	16
1	0.29	0.06	1.00	1.00	0.25	0.00	0.94	0.00	3	0	47	16

Table A.10 - SARS-CoV-2 1 thousand read sampling performance metrics. Thresholds were generated based on Bracken relative abundance measurements. ACC: Accuracy, TPR: True Positive Rate, TNR: True Negative Rate, PPV: Positive Predictive Value, NPV: Negative Predictive Value, FPR: False Positive Rate, FNR: False Negative Rate, FDR: False Discovery Rate, TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative.

Threshold	ACC	TPR	TNR	PPV	NPV	FPR	FNR	FDR	TP	FP	FN	TN
0	0.76	1.00	0.00	0.76	0.00	1.00	0.00	0.24	50	16	0	0
0.375	0.55	0.40	1.00	1.00	0.35	0.00	0.60	0.00	20	0	30	16
0.5	0.53	0.38	1.00	1.00	0.34	0.00	0.62	0.00	19	0	31	16
0.66055	0.52	0.36	1.00	1.00	0.33	0.00	0.64	0.00	18	0	32	16
1	0.50	0.34	1.00	1.00	0.33	0.00	0.66	0.00	17	0	33	16

Table A.11 - ICON DS Dry performance metrics for block size. Rows for hyperparameters with duplicate metrics to those preceding them (already represented in the below data table) were dropped (N=130 -> N=42). Hyperparameters with duplicate values were : [51, 55, 57, 83, 95, 97, 103, 105, 113, 115, 125, 129, 135, 137, 139, 141, 143, 145, 149, 153, 155, 157, 159, 161, 163, 169, 171, 173, 175, 177, 179, 181, 187, 189, 191, 193, 197, 199, 201, 203, 205, 207, 209, 211, 213, 215, 217, 219, 221, 223, 225, 227, 229, 231, 233, 235, 237, 239, 241, 243, 245, 247, 249, 251, 253, 255, 257, 259, 261, 263, 265, 267, 269, 271, 273, 275, 277, 279, 281, 283, 285, 287, 289, 291, 293, 295, 297, 299] . Hyp: Hyperparameter, ACC: Accuracy, TPR: True Positive Rate, TNR: True Negative Rate, PPV: Positive Predictive Value, NPV: Negative Predictive Value, FPR: False Positive Rate, FNR: False Negative Rate, FDR: False Discovery Rate, TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative.

Hyp	F1	AC C	TP R	TN R	PP V	NP V	FP R	FN R	FD R	TP	FP	FN	TN
41	0.66	0.61	0.51	0.92	0.95	0.38	0.08	0.49	0.05	73	4	71	44
43	0.65	0.61	0.49	0.96	0.97	0.39	0.04	0.51	0.03	71	2	73	46
45	0.65	0.61	0.48	1.00	1.00	0.39	0.00	0.52	0.00	69	0	75	48
47	0.67	0.63	0.50	1.00	1.00	0.40	0.00	0.50	0.00	72	0	72	48
49	0.66	0.62	0.49	1.00	1.00	0.40	0.00	0.51	0.00	71	0	73	48
53	0.65	0.61	0.49	1.00	1.00	0.39	0.00	0.51	0.00	70	0	74	48
59	0.68	0.64	0.52	1.00	1.00	0.41	0.00	0.48	0.00	75	0	69	48
61	0.71	0.66	0.55	1.00	1.00	0.42	0.00	0.45	0.00	79	0	65	48
63	0.70	0.66	0.54	1.00	1.00	0.42	0.00	0.46	0.00	78	0	66	48
65	0.73	0.68	0.57	1.00	1.00	0.44	0.00	0.43	0.00	82	0	62	48
67	0.73	0.68	0.58	1.00	1.00	0.44	0.00	0.42	0.00	83	0	61	48
69	0.75	0.70	0.60	1.00	1.00	0.46	0.00	0.40	0.00	87	0	57	48
71	0.78	0.73	0.65	1.00	1.00	0.48	0.00	0.35	0.00	93	0	51	48
73	0.79	0.74	0.65	1.00	1.00	0.49	0.00	0.35	0.00	94	0	50	48
75	0.82	0.78	0.70	1.00	1.00	0.53	0.00	0.30	0.00	101	0	43	48
77	0.85	0.80	0.74	1.00	1.00	0.56	0.00	0.26	0.00	106	0	38	48
79	0.86	0.81	0.75	1.00	1.00	0.57	0.00	0.25	0.00	108	0	36	48
81	0.87	0.83	0.77	1.00	1.00	0.59	0.00	0.23	0.00	111	0	33	48
85	0.87	0.82	0.76	1.00	1.00	0.59	0.00	0.24	0.00	110	0	34	48
87	0.88	0.84	0.79	1.00	1.00	0.62	0.00	0.21	0.00	114	0	30	48
89	0.88	0.84	0.78	1.00	1.00	0.61	0.00	0.22	0.00	113	0	31	48
91	0.89	0.85	0.80	1.00	1.00	0.62	0.00	0.20	0.00	115	0	29	48

93	0.89	0.85	0.81	1.00	1.00	0.63	0.00	0.19	0.00	116	0	28	48
99	0.90	0.86	0.81	1.00	1.00	0.64	0.00	0.19	0.00	117	0	27	48
101	0.90	0.86	0.82	1.00	1.00	0.65	0.00	0.18	0.00	118	0	26	48
107	0.90	0.87	0.83	1.00	1.00	0.66	0.00	0.17	0.00	119	0	25	48
109	0.91	0.88	0.83	1.00	1.00	0.67	0.00	0.17	0.00	120	0	24	48
111	0.92	0.89	0.85	1.00	1.00	0.69	0.00	0.15	0.00	122	0	22	48
117	0.93	0.90	0.86	1.00	1.00	0.71	0.00	0.14	0.00	124	0	20	48
119	0.93	0.90	0.87	1.00	1.00	0.72	0.00	0.13	0.00	125	0	19	48
121	0.93	0.91	0.88	1.00	1.00	0.73	0.00	0.13	0.00	126	0	18	48
123	0.94	0.91	0.88	1.00	1.00	0.74	0.00	0.12	0.00	127	0	17	48
127	0.94	0.92	0.89	1.00	1.00	0.75	0.00	0.11	0.00	128	0	16	48
131	0.95	0.93	0.91	1.00	1.00	0.79	0.00	0.09	0.00	131	0	13	48
133	0.96	0.94	0.92	1.00	1.00	0.80	0.00	0.08	0.00	132	0	12	48
147	0.96	0.94	0.92	1.00	1.00	0.81	0.00	0.08	0.00	133	0	11	48
151	0.96	0.95	0.93	1.00	1.00	0.83	0.00	0.07	0.00	134	0	10	48
165	0.97	0.95	0.94	1.00	1.00	0.84	0.00	0.06	0.00	135	0	9	48
167	0.97	0.96	0.94	1.00	1.00	0.86	0.00	0.06	0.00	136	0	8	48
183	0.98	0.96	0.95	1.00	1.00	0.87	0.00	0.05	0.00	137	0	7	48
185	0.98	0.97	0.96	1.00	1.00	0.89	0.00	0.04	0.00	138	0	6	48
195	0.98	0.97	0.97	1.00	1.00	0.91	0.00	0.03	0.00	139	0	5	48

Table A.12 - ICON DS Wet performance metrics for block size. Rows for hyperparameters with duplicate metrics to those preceding them (already represented in the below data table) were dropped (N=130 -> N=58). Hyperparameters with duplicate values were : [75, 97, 101, 103, 105, 119, 123, 125, 127, 129, 131, 137, 139, 145, 147, 149, 151, 153, 155, 157, 159, 161, 165, 171, 175, 177, 183, 185, 187, 189, 191, 193, 195, 197, 199, 201, 203, 205, 207, 209, 211, 213, 215, 217, 219, 223, 225, 227, 231, 233, 235, 237, 243, 245, 251, 253, 255, 257, 259, 261, 263, 265, 267, 271, 277, 281, 283, 285, 287, 291, 297, 299]. Hyp: Hyperparameter, ACC: Accuracy, TPR: True Positive Rate, TNR: True Negative Rate, PPV: Positive Predictive Value, NPV: Negative Predictive Value, FPR: False Positive Rate, FNR: False Negative Rate, FDR: False Discovery Rate, TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative.

Hyp	F1	ACC	TPR	TNR	PPV	NPV	FPR	FNR	FDR	TP	FP	FN	TN
41	0.75	0.72	0.67	0.82	0.86	0.60	0.18	0.33	0.14	80	13	40	59
43	0.74	0.71	0.64	0.83	0.87	0.58	0.17	0.36	0.13	77	12	43	60
45	0.73	0.71	0.63	0.86	0.88	0.58	0.14	0.38	0.12	75	10	45	62
47	0.73	0.72	0.61	0.92	0.92	0.58	0.08	0.39	0.08	73	6	47	66
49	0.75	0.73	0.63	0.92	0.93	0.59	0.08	0.38	0.07	75	6	45	66
51	0.74	0.73	0.61	0.94	0.95	0.59	0.06	0.39	0.05	73	4	47	68
53	0.75	0.74	0.61	0.97	0.97	0.60	0.03	0.39	0.03	73	2	47	70
55	0.75	0.74	0.60	0.99	0.99	0.60	0.01	0.40	0.01	72	1	48	71
57	0.74	0.74	0.59	0.99	0.99	0.59	0.01	0.41	0.01	71	1	49	71
59	0.68	0.70	0.53	0.99	0.98	0.55	0.01	0.48	0.02	63	1	57	71
61	0.72	0.72	0.56	1.00	1.00	0.58	0.00	0.44	0.00	67	0	53	72
63	0.72	0.73	0.57	1.00	1.00	0.58	0.00	0.43	0.00	68	0	52	72
65	0.74	0.74	0.58	1.00	1.00	0.59	0.00	0.42	0.00	70	0	50	72
67	0.78	0.78	0.64	1.00	1.00	0.63	0.00	0.36	0.00	77	0	43	72
69	0.77	0.77	0.63	1.00	1.00	0.62	0.00	0.38	0.00	75	0	45	72
71	0.80	0.79	0.67	1.00	1.00	0.64	0.00	0.33	0.00	80	0	40	72
73	0.81	0.80	0.68	1.00	1.00	0.65	0.00	0.32	0.00	82	0	38	72
77	0.81	0.80	0.68	1.00	1.00	0.65	0.00	0.33	0.00	81	0	39	72
79	0.82	0.81	0.70	1.00	1.00	0.67	0.00	0.30	0.00	84	0	36	72
81	0.83	0.82	0.71	1.00	1.00	0.67	0.00	0.29	0.00	85	0	35	72
83	0.84	0.83	0.73	1.00	1.00	0.69	0.00	0.28	0.00	87	0	33	72
85	0.85	0.84	0.74	1.00	1.00	0.70	0.00	0.26	0.00	89	0	31	72
87	0.86	0.84	0.75	1.00	1.00	0.71	0.00	0.25	0.00	90	0	30	72
89	0.88	0.87	0.79	1.00	1.00	0.74	0.00	0.21	0.00	95	0	25	72

91	0.90	0.89	0.82	1.00	1.00	0.77	0.00	0.18	0.00	98	0	22	72
93	0.89	0.88	0.81	1.00	1.00	0.76	0.00	0.19	0.00	97	0	23	72
95	0.92	0.91	0.85	1.00	1.00	0.80	0.00	0.15	0.00	102	0	18	72
99	0.93	0.92	0.88	1.00	1.00	0.83	0.00	0.13	0.00	105	0	15	72
107	0.94	0.93	0.88	1.00	1.00	0.84	0.00	0.12	0.00	106	0	14	72
109	0.94	0.93	0.89	1.00	1.00	0.85	0.00	0.11	0.00	107	0	13	72
111	0.95	0.94	0.91	1.00	1.00	0.87	0.00	0.09	0.00	109	0	11	72
113	0.96	0.95	0.93	1.00	1.00	0.89	0.00	0.08	0.00	111	0	9	72
115	0.97	0.96	0.93	1.00	1.00	0.90	0.00	0.07	0.00	112	0	8	72
117	0.97	0.96	0.94	1.00	1.00	0.91	0.00	0.06	0.00	113	0	7	72
121	0.98	0.97	0.96	1.00	1.00	0.94	0.00	0.04	0.00	115	0	5	72
133	0.97	0.97	0.96	0.99	0.99	0.93	0.01	0.04	0.01	115	1	5	71
135	0.98	0.97	0.97	0.99	0.99	0.95	0.01	0.03	0.01	116	1	4	71
141	0.97	0.97	0.97	0.97	0.98	0.95	0.03	0.03	0.02	116	2	4	70
143	0.98	0.97	0.98	0.97	0.98	0.96	0.03	0.03	0.02	117	2	3	70
163	0.98	0.97	0.98	0.96	0.98	0.96	0.04	0.03	0.03	117	3	3	69
167	0.98	0.98	0.99	0.96	0.98	0.99	0.04	0.01	0.02	119	3	1	69
169	0.99	0.98	1.00	0.96	0.98	1.00	0.04	0.00	0.02	120	3	0	69
173	0.98	0.98	1.00	0.94	0.97	1.00	0.06	0.00	0.03	120	4	0	68
179	0.98	0.97	0.99	0.94	0.97	0.99	0.06	0.01	0.03	119	4	1	68
181	0.98	0.97	0.98	0.94	0.97	0.97	0.06	0.02	0.03	118	4	2	68
221	0.98	0.97	0.99	0.93	0.96	0.99	0.07	0.01	0.04	119	5	1	67
229	0.97	0.96	0.99	0.92	0.95	0.99	0.08	0.01	0.05	119	6	1	66
239	0.97	0.96	0.98	0.92	0.95	0.97	0.08	0.02	0.05	118	6	2	66
241	0.96	0.95	0.98	0.90	0.94	0.97	0.10	0.02	0.06	118	7	2	65
247	0.96	0.95	0.98	0.89	0.94	0.97	0.11	0.02	0.06	118	8	2	64
249	0.96	0.94	0.98	0.88	0.93	0.97	0.13	0.02	0.07	118	9	2	63
269	0.95	0.94	0.98	0.86	0.92	0.97	0.14	0.02	0.08	118	10	2	62
273	0.95	0.93	0.98	0.85	0.91	0.97	0.15	0.02	0.09	118	11	2	61
275	0.94	0.93	0.98	0.83	0.91	0.97	0.17	0.02	0.09	118	12	2	60
279	0.94	0.92	0.98	0.83	0.91	0.95	0.17	0.03	0.09	117	12	3	60
289	0.94	0.92	0.98	0.82	0.90	0.95	0.18	0.03	0.10	117	13	3	59
293	0.93	0.91	0.98	0.81	0.89	0.95	0.19	0.03	0.11	117	14	3	58
295	0.93	0.91	0.98	0.79	0.89	0.95	0.21	0.03	0.11	117	15	3	57

Table A.13 - QUIDEL Dry performance metrics for block size. Rows for hyperparameters with duplicate metrics to those preceding them (already represented in the below data table) were dropped (N=130 -> N=32). Hyperparameters with duplicate values were : [71, 75, 77, 79, 81, 83, 85, 87, 89, 91, 99, 101, 103, 105, 107, 109, 119, 121, 123, 125, 127, 129, 131, 135, 137, 139, 141, 143, 145, 147, 149, 153, 157, 165, 167, 169, 171, 173, 175, 177, 179, 181, 183, 185, 187, 189, 191, 195, 197, 199, 201, 203, 205, 207, 209, 211, 213, 215, 217, 219, 221, 223, 225, 229, 231, 233, 235, 237, 239, 241, 243, 245, 247, 249, 251, 255, 257, 259, 261, 263, 265, 267, 269, 271, 273, 275, 277, 279, 281, 283, 285, 287, 289, 291, 293, 295, 297, 299]. Hyp: Hyperparameter, ACC: Accuracy, TPR: True Positive Rate, TNR: True Negative Rate, PPV: Positive Predictive Value, NPV: Negative Predictive Value, FPR: False Positive Rate, FNR: False Negative Rate, FDR: False Discovery Rate, TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative.

Hyp	F1	ACC	TPR	TNR	PPV	NPV	FPR	FNR	FDR	TP	FP	FN	TN
41	0.79	0.70	0.65	1.00	1.00	0.29	0.00	0.35	0.00	110	0	58	24
43	0.82	0.73	0.70	1.00	1.00	0.32	0.00	0.30	0.00	117	0	51	24
45	0.83	0.74	0.71	1.00	1.00	0.33	0.00	0.29	0.00	119	0	49	24
47	0.83	0.75	0.71	1.00	1.00	0.33	0.00	0.29	0.00	120	0	48	24
49	0.84	0.76	0.72	1.00	1.00	0.34	0.00	0.28	0.00	121	0	47	24
51	0.85	0.77	0.73	1.00	1.00	0.35	0.00	0.27	0.00	123	0	45	24
53	0.86	0.78	0.75	1.00	1.00	0.36	0.00	0.25	0.00	126	0	42	24
55	0.87	0.80	0.77	1.00	1.00	0.39	0.00	0.23	0.00	130	0	38	24
57	0.88	0.81	0.78	1.00	1.00	0.39	0.00	0.22	0.00	131	0	37	24
59	0.89	0.82	0.80	1.00	1.00	0.41	0.00	0.20	0.00	134	0	34	24
61	0.89	0.83	0.81	1.00	1.00	0.43	0.00	0.19	0.00	136	0	32	24
63	0.90	0.84	0.82	1.00	1.00	0.44	0.00	0.18	0.00	138	0	30	24
65	0.91	0.86	0.84	1.00	1.00	0.47	0.00	0.16	0.00	141	0	27	24
67	0.92	0.86	0.85	1.00	1.00	0.48	0.00	0.15	0.00	142	0	26	24
69	0.92	0.87	0.85	1.00	1.00	0.49	0.00	0.15	0.00	143	0	25	24
73	0.92	0.88	0.86	1.00	1.00	0.50	0.00	0.14	0.00	144	0	24	24
93	0.93	0.88	0.86	1.00	1.00	0.51	0.00	0.14	0.00	145	0	23	24
95	0.93	0.89	0.87	1.00	1.00	0.52	0.00	0.13	0.00	146	0	22	24
97	0.93	0.89	0.88	1.00	1.00	0.53	0.00	0.13	0.00	147	0	21	24
111	0.94	0.90	0.88	1.00	1.00	0.55	0.00	0.12	0.00	148	0	20	24
113	0.94	0.91	0.89	1.00	1.00	0.57	0.00	0.11	0.00	150	0	18	24
115	0.95	0.92	0.90	1.00	1.00	0.60	0.00	0.10	0.00	152	0	16	24
117	0.95	0.91	0.90	1.00	1.00	0.59	0.00	0.10	0.00	151	0	17	24
133	0.95	0.92	0.91	1.00	1.00	0.62	0.00	0.09	0.00	153	0	15	24

151	0.96	0.93	0.92	1.00	1.00	0.63	0.00	0.08	0.00	154	0	14	24
155	0.96	0.94	0.93	1.00	1.00	0.67	0.00	0.07	0.00	156	0	12	24
159	0.97	0.94	0.93	1.00	1.00	0.69	0.00	0.07	0.00	157	0	11	24
161	0.97	0.95	0.94	1.00	1.00	0.71	0.00	0.06	0.00	158	0	10	24
163	0.97	0.95	0.95	1.00	1.00	0.73	0.00	0.05	0.00	159	0	9	24
193	0.98	0.96	0.95	1.00	1.00	0.75	0.00	0.05	0.00	160	0	8	24
227	0.96	0.93	0.92	1.00	1.00	0.65	0.00	0.08	0.00	155	0	13	24
253	0.94	0.90	0.89	1.00	1.00	0.56	0.00	0.11	0.00	149	0	19	24

Table A.14 - QUIDEL Wet performance metrics for block size. Rows for hyperparameters with duplicate metrics to those preceding them (already represented in the below data table) were dropped (N=130 -> N=41). Hyperparameters with duplicate values were : [57, 67, 91, 93, 95, 97, 99, 103, 105, 109, 121, 123, 125, 129, 131, 133, 135, 147, 149, 151, 155, 157, 159, 163, 165, 167, 169, 177, 179, 181, 183, 185, 187, 189, 191, 193, 195, 197, 199, 201, 203, 205, 207, 209, 211, 213, 215, 217, 219, 221, 223, 225, 227, 229, 231, 233, 235, 237, 239, 241, 243, 245, 247, 249, 251, 253, 255, 257, 259, 261, 263, 265, 267, 269, 271, 273, 275, 277, 279, 281, 283, 285, 287, 289, 291, 293, 295, 297, 299]. Hyp: Hyperparameter, ACC: Accuracy, TPR: True Positive Rate, TNR: True Negative Rate, PPV: Positive Predictive Value, NPV: Negative Predictive Value, FPR: False Positive Rate, FNR: False Negative Rate, FDR: False Discovery Rate, TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative.

Hyp	F1	ACC	TPR	TNR	PPV	NPV	FPR	FNR	FDR	TP	FP	FN	TN
41	0.65	0.61	0.49	1.00	1.00	0.39	0.00	0.51	0.00	70	0	74	48
43	0.67	0.63	0.51	1.00	1.00	0.40	0.00	0.49	0.00	73	0	71	48
45	0.68	0.64	0.52	1.00	1.00	0.41	0.00	0.48	0.00	75	0	69	48
47	0.71	0.66	0.55	1.00	1.00	0.42	0.00	0.45	0.00	79	0	65	48
49	0.73	0.68	0.57	1.00	1.00	0.44	0.00	0.43	0.00	82	0	62	48
51	0.74	0.69	0.59	1.00	1.00	0.45	0.00	0.41	0.00	85	0	59	48
53	0.76	0.71	0.62	1.00	1.00	0.47	0.00	0.38	0.00	89	0	55	48
55	0.77	0.72	0.63	1.00	1.00	0.48	0.00	0.37	0.00	91	0	53	48
59	0.78	0.73	0.64	1.00	1.00	0.48	0.00	0.36	0.00	92	0	52	48
61	0.79	0.74	0.65	1.00	1.00	0.49	0.00	0.35	0.00	94	0	50	48
63	0.79	0.74	0.66	1.00	1.00	0.49	0.00	0.34	0.00	95	0	49	48
65	0.81	0.76	0.68	1.00	1.00	0.51	0.00	0.32	0.00	98	0	46	48
69	0.82	0.77	0.69	1.00	1.00	0.52	0.00	0.31	0.00	100	0	44	48
71	0.83	0.78	0.71	1.00	1.00	0.53	0.00	0.29	0.00	102	0	42	48
73	0.84	0.80	0.73	1.00	1.00	0.55	0.00	0.27	0.00	105	0	39	48
75	0.86	0.81	0.75	1.00	1.00	0.57	0.00	0.25	0.00	108	0	36	48
77	0.86	0.82	0.76	1.00	1.00	0.58	0.00	0.24	0.00	109	0	35	48
79	0.87	0.82	0.76	1.00	1.00	0.59	0.00	0.24	0.00	110	0	34	48
81	0.87	0.83	0.77	1.00	1.00	0.59	0.00	0.23	0.00	111	0	33	48
83	0.89	0.85	0.80	1.00	1.00	0.62	0.00	0.20	0.00	115	0	29	48
85	0.89	0.85	0.81	1.00	1.00	0.63	0.00	0.19	0.00	116	0	28	48
87	0.90	0.86	0.82	1.00	1.00	0.65	0.00	0.18	0.00	118	0	26	48
89	0.90	0.87	0.83	1.00	1.00	0.66	0.00	0.17	0.00	119	0	25	48
101	0.91	0.88	0.83	1.00	1.00	0.67	0.00	0.17	0.00	120	0	24	48
107	0.91	0.88	0.84	1.00	1.00	0.68	0.00	0.16	0.00	121	0	23	48
111	0.92	0.89	0.85	1.00	1.00	0.69	0.00	0.15	0.00	122	0	22	48
113	0.92	0.89	0.85	1.00	1.00	0.70	0.00	0.15	0.00	123	0	21	48
115	0.93	0.90	0.87	1.00	1.00	0.72	0.00	0.13	0.00	125	0	19	48

117	0.93	0.91	0.88	1.00	1.00	0.73	0.00	0.13	0.00	126	0	18	48
119	0.94	0.91	0.88	1.00	1.00	0.74	0.00	0.12	0.00	127	0	17	48
127	0.94	0.92	0.89	1.00	1.00	0.75	0.00	0.11	0.00	128	0	16	48
137	0.95	0.92	0.90	1.00	1.00	0.76	0.00	0.10	0.00	129	0	15	48
139	0.95	0.93	0.90	1.00	1.00	0.77	0.00	0.10	0.00	130	0	14	48
141	0.95	0.93	0.91	1.00	1.00	0.79	0.00	0.09	0.00	131	0	13	48
143	0.96	0.94	0.92	1.00	1.00	0.80	0.00	0.08	0.00	132	0	12	48
145	0.96	0.94	0.92	1.00	1.00	0.81	0.00	0.08	0.00	133	0	11	48
153	0.96	0.95	0.93	1.00	1.00	0.83	0.00	0.07	0.00	134	0	10	48
161	0.97	0.95	0.94	1.00	1.00	0.84	0.00	0.06	0.00	135	0	9	48
171	0.97	0.96	0.94	1.00	1.00	0.86	0.00	0.06	0.00	136	0	8	48
173	0.98	0.96	0.95	1.00	1.00	0.87	0.00	0.05	0.00	137	0	7	48
175	0.98	0.97	0.96	1.00	1.00	0.89	0.00	0.04	0.00	138	0	6	48