Journal Homepage: -www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI:10.21474/IJAR01/19548
DOI URL: http://dx.doi.org/10.21474/IJAR01/19548

RESEARCH ARTICLE

PATHOEXTRACT: A BIOINFORMATIC PIPELINE FOR QUALITY CONTROL AND HOST DNA REMOVAL IN PLASMODIUM FALCIPARUM NGS DATA

Stanislas Egomli Assohoun[1,2,3,4], Aristide Berenger Ako[2,3], Patrice N'guessan Akoguhi[4], Paul Christian Abouchou Ako[3], Medard Brou Kouassi[1], Jérôme Adou Kablan[1] and Ronan Jambou[2]

1. Laboratory of Mechanics and Computer Science (LAMI), Félix Houphouët-Boigny University, Abidjan, Côte d'Ivoire.
2. Parasitology and Mycology Unit, Pasteur Institute (IPCI), Abidjan, Côte d'Ivoire.
3. Genomics and Metagenomics Platform, Pasteur Institute (IPCI), Abidjan, Côte d'Ivoire.
4. Laboratory of Environmental Science and Technology (LSTE), Jean LorougnonGuédé University (UJLoG), Daloa, Côte d'Ivoire.

Manuscript Info

Manuscript History
Received: 28 July 2024
Final Accepted: 30 August 2024
Published: September 2024

Key words:-
Next-Generation Sequencing, Decontamination, Quality Control, Plasmodium Falciparum, Bioinformatics

Abstract

Malaria, caused by Plasmodium falciparum, is a significant global health burden, particularly in sub-Saharan Africa. Deep sequencing (NGS) of parasite genomes has revolutionized our understanding of its biology and the emergence of drug resistance. However, the presence of host human DNA and other microbial contaminants within patient samples can hinder accurate and efficient parasite genome analysis. To address this challenge, we have developed PathoExtract, a robust bioinformatics pipeline that integrates commonly used tools into a streamlined workflow. PathoExtract leverages Snakemake, a workflow management system, to provide a flexible and reproducible framework for data processing. The pipeline incorporates rigorous quality control steps to identify and remove low-quality reads and contaminants. Host DNA and microbial sequences are effectively filtered out using a combination of alignment-based and alignment-free methods, ensuring that only Plasmodium falciparum reads are retained for downstream analysis.The pipeline offers an intuitive graphical user interface, making it accessible to researchers with varying levels of bioinformatics expertise. This user-friendly interface simplifies the process of running the pipeline, even for those unfamiliar with command-line tools. The code and documentation for PathoExtract are freely available at: https://github.com/stanlasso/DREPAL-PATHOEXTRACT.

Introduction:-

Malaria, caused by the Plasmodium falciparum parasite, remains one of the leading causes of morbidity and mortality in tropical and subtropical regions, particularly in sub-Saharan Africa. This infectious disease affects millions of people each year, resulting in hundreds of thousands of deaths, predominantly among children under five and pregnant women[1]. While significant progress has been made in reducing malaria transmission through

Corresponding Author: Stanislas Egomli Assohoun
Address:- Jean Lorougnon Guédé University (UJLoG), BP 150, Daloa, Côte d'Ivoire.

interventions such as insecticide-treated nets and artemisinin-based combination therapies, the emergence of drug resistance, especially artemisinin resistance, poses a serious challenge to achieving global malaria control targets.

Drug resistance in Plasmodium falciparum has been linked to specific genetic mutations, which can compromise the efficacy of first-line treatments. Continuous surveillance of these mutations is essential for adapting therapeutic strategies and ensuring the effectiveness of control programs [2]. Genomic studies provide insights into the evolution of resistance, enabling researchers to detect and monitor resistant strains across different regions. Advances in next-generation sequencing (NGS) have revolutionized our ability to perform in-depth analyses of the parasite genome, uncovering a wide array of genetic variations that may influence treatment outcomes. However, one of the significant challenges in utilizing NGS data from clinical samples is contamination by host human DNA and microorganisms, which complicates the identification of parasite sequences and risks leading to inaccuracies in genomic analyses [3].

The complexity of handling contaminated NGS data requires sophisticated bioinformatics approaches to ensure the reliable extraction of parasite-specific sequences. Several tools, including Fastqc(https://github.com/s-andrews/FastQC/blob/master/fastqc), Multiqc[4], Fastp[5], Trimgalore(https://github.com/FelixKrueger/TrimGalore BWA [6], and Samtools [7], have been developed to process raw NGS data, but their integration into coherent workflows remains difficult for many researchers, particularly those without bioinformatics expertise. Existing pipelines like HoCoRT[8]DeconSeq[9], and Fastq Screen [10] offer partial solutions but often lack flexibility or a user-friendly interface, making them inaccessible to a broader scientific audience. Therefore, there is a need for a streamlined tool that can integrate these functionalities into an accessible and reproducible pipeline, enabling more accurate and efficient data analysis.

To address these challenges, we developed PathoExtract, a bioinformatics tool specifically designed for the quality control and decontamination of NGS data from Plasmodium falciparum-infected samples. This tool was developed as part of the DREPAL project (Sickle Cell Disease-Malaria), led by the Department of Parasitology-Mycology at the Institut Pasteur of Côte d'Ivoire (IPCI). PathoExtract integrates widely used bioinformatics software, such as Fastqc, Multiqc, Trimgalore, Megahit[11] BWA, and Samtools, into a modular pipeline using Snakemake[12], a workflow management system that simplifies the automation of complex analyses. This integration allows for the efficient removal of human host DNA and microbial contaminants, thereby facilitating the isolation and analysis of the parasite genome.

PathoExtract offers an intuitive graphical interface that simplifies the user experience, even for researchers without advanced bioinformatics training. By automating the steps required for NGS data cleaning and providing clear, visualized outputs, the tool democratizes access to high-quality genomic data analysis. The ability to efficiently remove contaminants and focus on Plasmodium falciparum sequences significantly enhances the reliability of downstream analyses, such as the detection of drug resistance mutations. This, in turn, provides crucial data for guiding malaria treatment strategies and improving public health interventions.

Beyond its technical capabilities, PathoExtract contributes to the broader effort of monitoring malaria resistance at the genomic level. By offering an accessible and scalable solution, the tool supports the identification of key mutations, allowing researchers to track Plasmodium falciparum variants across different geographic regions. The insights gained from these analyses are essential for adapting malaria control strategies in real time, ensuring their continued effectiveness as parasite populations evolve. PathoExtract thus represents a major advancement in NGS data preprocessing, facilitating both clinical and epidemiological research on malaria.

Material and Method:-

PathoExtract Architecture.

PathoExtract is a bioinformatics tool designed for the quality control and decontamination of NGS data, specifically targeting samples from Plasmodium falciparum-positive patients. Its architecture (Figure 1) features an intuitive front-end interface built with AngularJS (https://github.com/angular/angular.js/), making it accessible to researchers without bioinformatics expertise. The backend server is powered by NodeJS(https://github.com/nodejs/node) with PM2 (https://github.com/Unitech/pm2) used as the process manager. The orchestration of the different PathoExtract pipelines relies on snakemake, a modular and automated workflow manager, to simplify data processing. It integrates several open-source tools widely adopted by the NGS community for the analysis of raw data.

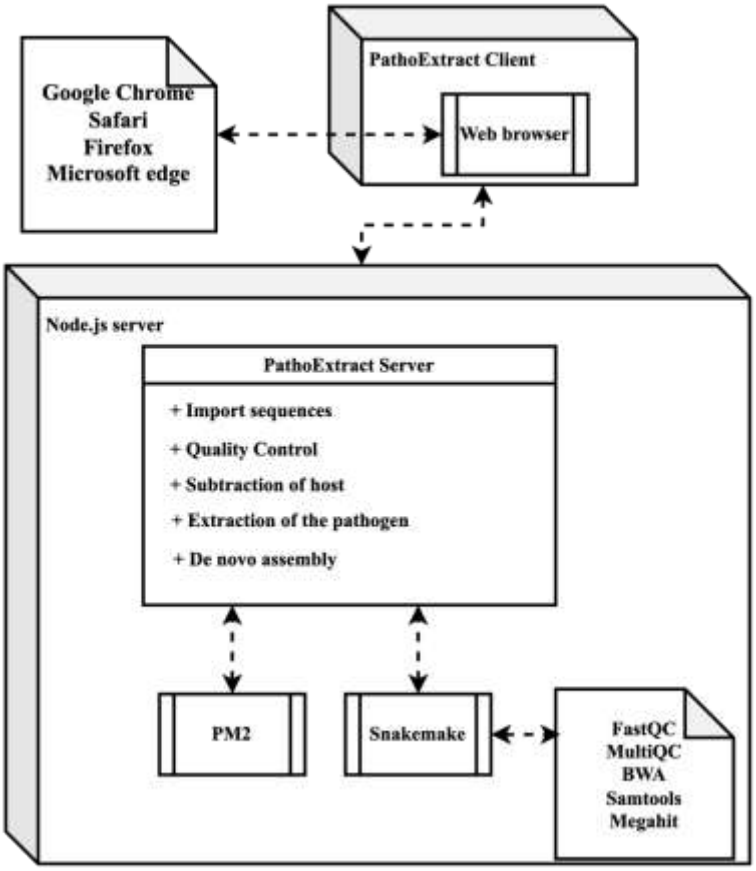

Figure 1:- PathoExtract Architecture.

General workflow of PathoExtract.

PathoExtract relies on an automated processing pipeline that integrates several essential bioinformatics steps (Figure 2) :

Quality control (QC) :

The initial stage of the analysis pipeline involves rigorous quality control (QC) measures. Tools like Fastqc and Multiqc provide a comprehensive assessment of raw sequencing data, identifying potential artifacts and low-quality sequences. This empowers researchers to make informed decisions about data inclusion for downstream analyses. Subsequently, based on predetermined quality parameters, Trimgalore performs read adaptation. This process entails selective removal of reads or read portions that fall outside user-defined quality thresholds. Specifically, Trimgalore eliminates low-quality bases flanking the 5' and 3' ends of reads. Additionally, reads containing an excessive number of ambiguous nucleotides ("Ns") are discarded. Finally, reads failing to meet minimum length or average quality score cutoffs are removed

Digitalfiltering (decontamination and extraction of parasite sequences) :

Digital filtering of raw sequencing data involves a two-tiered alignment process to eliminate host and microbial contaminants. Reads are initially aligned to the human reference genome (GRCh38) using BWA, and unaligned reads are subsequently aligned to the Plasmodium falciparum reference genome (3D7). This approach effectively removes human DNA sequences and other potential contaminants, ensuring the purity of the parasite-derived data. Samtools is then employed to extract mapped and unmapped reads, and to convert BAM files into fastq format for downstream analysis. The resulting outputs include reads aligned to the human genome (host mapped), reads not aligned to the human genome (host unmapped), reads aligned to the Plasmodium falciparum genome (pathogen mapped), and remaining unaligned reads (others). This step is critical for obtaining high-quality, pathogen-specific sequencing data that is suitable for further bioinformatic analyses.

De novo assembly:
After the digital filtering step, the parasitic sequences are assembled de novo using Megahit, coupled with QUAST for assembly visualization.

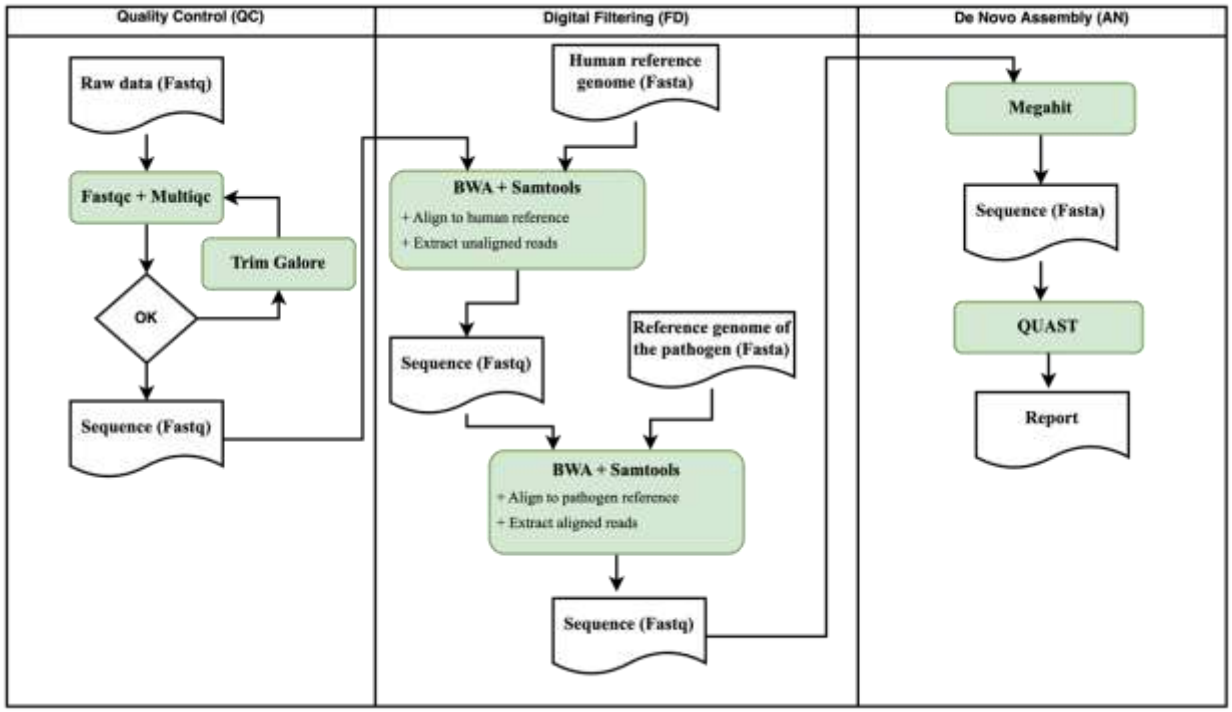

Figure 2:- Detailed diagram of the overall workflow of the PathoExtract analysis pipeline.

Datasets. PathoExtract was tested on two NGS datasets :

1. Fifteen paired-end read samples from an Illumina NGS whole-genome sequencing of Plasmodium falciparum genomes, derived from plasmodial DNA obtained from the continuous in vitro culture (parasite maturation) of whole blood from malaria patients. These data are from the DREPAL (Sickle Cell Disease – Malaria) project, which took place at the Pasteur Institute of Côte d'Ivoire (IPCI), specifically in the Parasitology laboratory.
2. Fifteen paired-end read samples from whole-genome sequencing on the Illumina HiSeq 2000 platform, obtained from the European Nucleotide Archive (ENA).

Hardware and Software Environment.
The analyses were performed on an HPE ProLiant DL380 Gen9 server equipped with an Intel Xeon E5-2620V3 processor, 64 GB of RAM, and two hard drives (1 TB and 7 TB). The operating system used was Linux Ubuntu.

Results and Discussion:-

Results:-

Graphical User Interface (GUI).
The PathoExtract graphical interface has been designed to be intuitive and accessible to researchers, even those without advanced technical expertise. It allows users to easily configure and execute the various modules of the tool, while clearly displaying results and reports. The interface offers options for sample loading, treatment parameter selection, as well as access to execution logs and analysis outcomes. The main screens of this interface are illustrated in Figure 3. Figure 3.A and Figure 3.B depict the starting points of the primary analysis workflows of PathoExtract. They respectively show the options for loading or indexing, if necessary, the reference genomes of the host and the pathogen, along with uploading patient samples into the application's analysis directory. Figure 3.C presents the various parameters provided by the user to activate the digital filtering pipeline, which primarily includes the host and pathogen reference genomes used in this analysis. Lastly, Figure 3.D outlines the sequential steps leading to the assembly of reads.

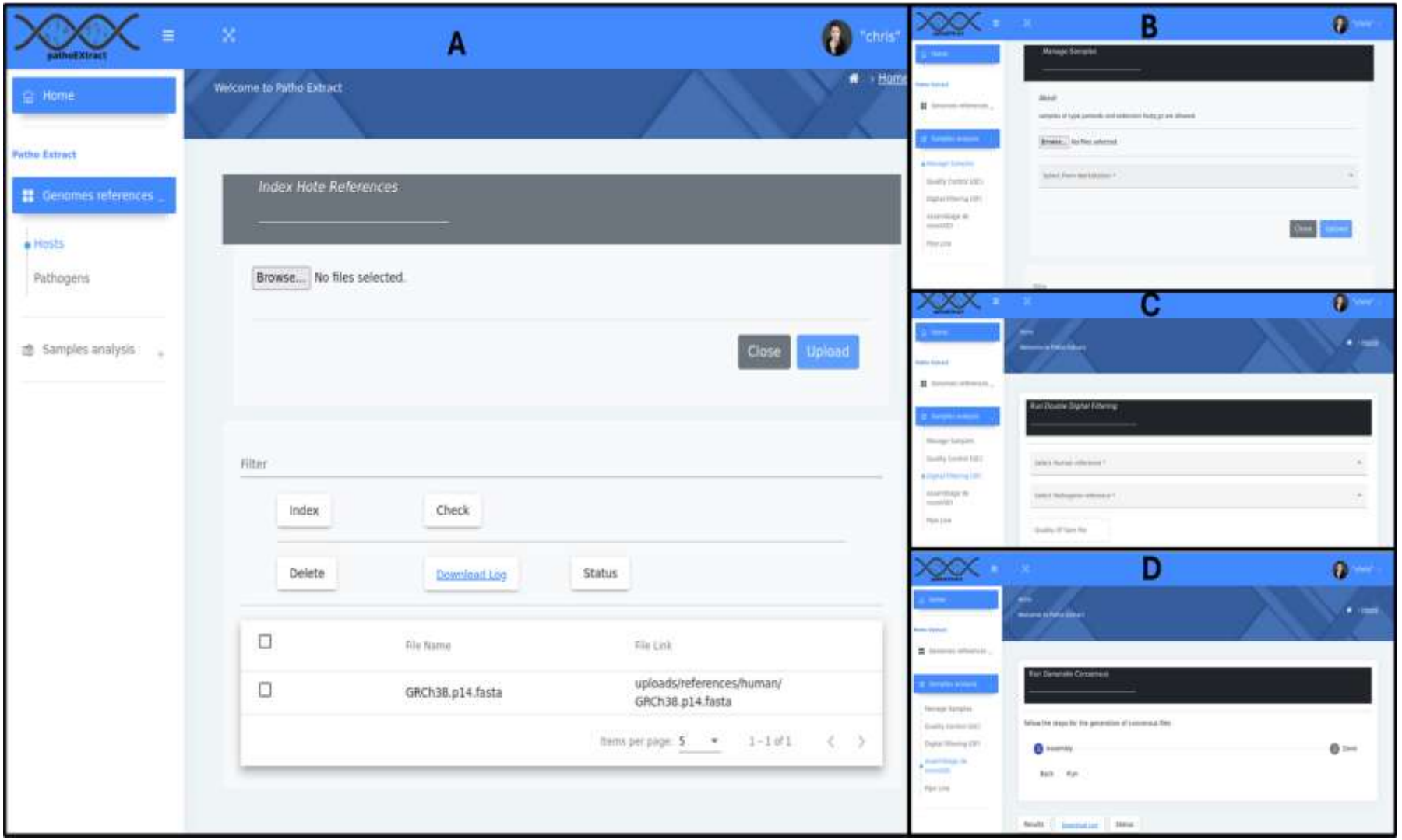

Figure 3:- Main screenshots of PathoExtract.

Testing and Validation of PathoExtract on Biological Data.

PathoExtract was evaluated through three key modules: quality control, digital filtering, and de novo assembly, each crucial for optimizing NGS data from Plasmodium falciparum-infected samples. Two experiments were conducted to assess its effectiveness in improving data quality, decontamination, and the extraction of parasite-specific reads.

Experiment 1.

The qualitycontrol module of PathoExtract processed the NGS data samples in this study using Fastqc and Multiqc to visualize and assess the quality of the reads. Data quality was optimized by removing low-quality reads and undetermined bases (Ns). After this step, the samples exhibited a quality distribution consistent with the defined parameters, with a notable reduction in inappropriate reads, ensuring that only high-quality reads were retained for subsequent steps. During the Digital Filtering step, PathoExtract removes host (human) contaminants before extracting parasite reads. Figure 1 presents the statistics for the digital filtering (DF) performed. It shows the proportions of reads: (i) related to the human host (Host), (ii) the pathogen of interest (Pathogen), and (iii) other unidentified sources (Other).

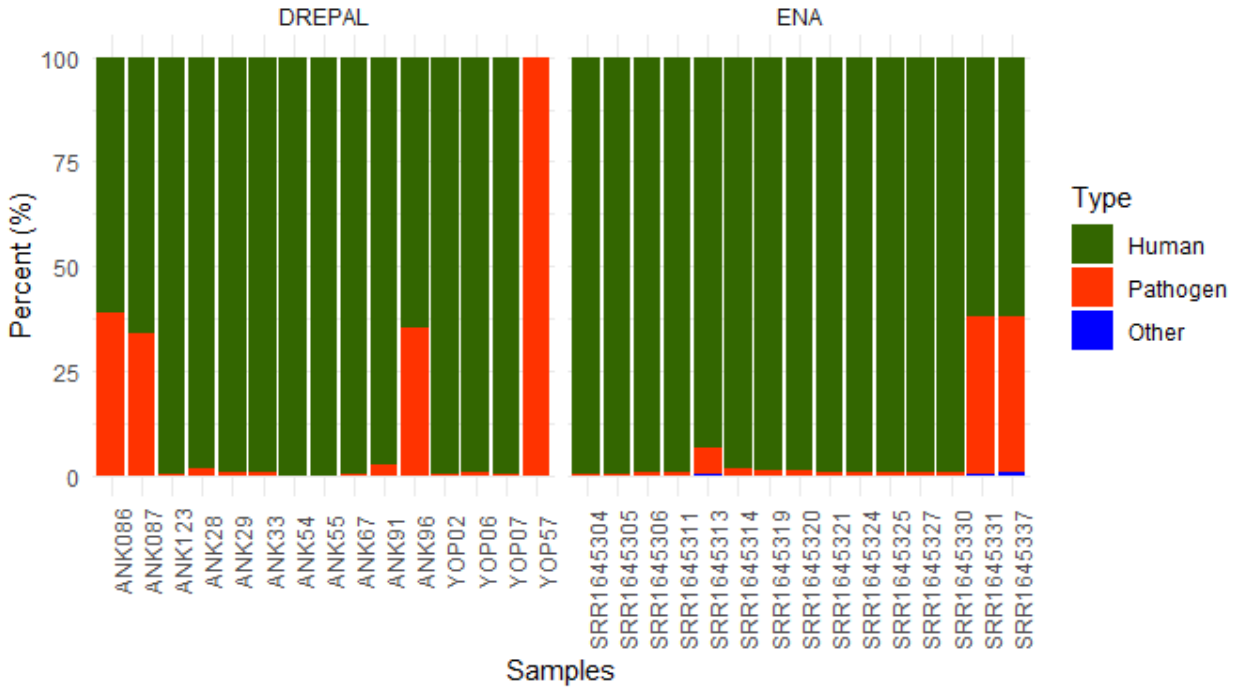

Figure 4:- Statistics obtained fron digital filtering.

With the exception of sample YOP57, which showed no reads aligned to the human genome, the proportion of human-derived reads ranged from 61.16% to 99.98% for the DREPAL project samples, compared to 62.14% to 99.73% for the ENA database samples. Regarding the reads aligned to the pathogen of interest, Plasmodium falciparum, the observed rates ranged from 0.04% to 38.84% for the DREPAL samples, and from 0.17% to 37.41% for the ENA samples, with the exception of sample YOP57, which exhibited 100% parasitic alignment. Finally, the proportion of reads classified as "other sources" was low, with values ranging from 0% to 0.69% across all analyzed samples.

Experiment 2:

To further verify the quality (specificity of the results) obtained with PathoExtract during the first experiment, we also applied FastQ Screen in addition to PathoExtract to check the level of residual contamination in the outputs that were unaligned with the human genome (Host unmapped) and those aligned with Plasmodium falciparum (Patho mapped) from the DREPAL project data. The results of this second experiment are summarized in Tables 1 and 2.

Table 1:- Human and parasitic (Pf) reads detected by PathoExtract and FastQ Screen in the unmapped human output from the first experiment.

			PathoExtract				**Fastq Screen**			
			Number of reads / % of reads							
Sample	**R1/R2**	**Total**	**Humain**	**%**	**Pf**	**%**	**Humain**	**%**	**Pf**	**%**
ANK086	R1	7542661	0	**0**	5797080	**76,86**	0	**0**	5844959	**77,49**
	R2	7542661	0	**0**	5797080	**76,86**	0	**0**	5794111	**76,82**
ANK087	R1	5089094	0	**0**	4440445	**87,25**	0	**0**	4482404	**88,08**
	R2	5089094	0	**0**	4440445	**87,25**	0	**0**	4434353	**87,13**
ANK123	R1	2109196	0	**0**	34568	**1,64**	0	**0**	35094	**1,66**
	R2	2109196	0	**0**	34568	**1,64**	0	**0**	34661	**1,64**
ANK28	R1	168627	0	**0**	158078	**93,74**	0	**0**	158035	**93,72**
	R2	168627	0	**0**	158078	**93,74**	0	**0**	157088	**93,16**
ANK29	R1	170419	0	**0**	151485	**88,89**	0	**0**	151167	**88,70**
	R2	170419	0	**0**	151485	**88,89**	0	**0**	150036	**88,04**
ANK33	R1	405131	0	**0**	106877	**26,38**	0	**0**	107078	**26,43**
	R2	405131	0	**0**	106877	**26,38**	0	**0**	105367	**26,01**
ANK54	R1	11922	0	**0**	6960	**58,38**	0	**0**	6891	**57,80**
	R2	11922	0	**0**	6960	**58,38**	0	**0**	6837	**57,35**
ANK55	R1	14037	0	**0**	4317	**30,75**	0	**0**	4264	**30,38**
	R2	14037	0	**0**	4317	**30,75**	0	**0**	4236	**30,18**
ANK67	R1	74914	0	**0**	47576	**63,51**	0	**0**	47589	**63,52**
	R2	74914	0	**0**	47576	**63,51**	0	**0**	46912	**62,62**
ANK91	R1	441988	0	**0**	210450	**47,61**	0	**0**	211734	**47,90**
	R2	441988	0	**0**	210450	**47,61**	0	**0**	210893	**47,71**
ANK96	R1	6603373	0	**0**	6431181	**97,39**	0	**0**	6441507	**97,55**
	R2	6603373	0	**0**	6431181	**97,39**	0	**0**	6393483	**96,82**
YOP02	R1	36540	0	**0**	31455	**86,08**	0	**0**	31422	**85,99**
	R2	36540	0	**0**	31455	**86,08**	0	**0**	31212	**85,42**
YOP06	R1	138531	0	**0**	125959	**90,92**	0	**0**	125873	**90,86**
	R2	138531	0	**0**	125959	**90,92**	0	**0**	124836	**90,11**
YOP07	R1	62859	0	**0**	46164	**73,44**	0	**0**	46104	**73,35**
	R2	62859	0	**0**	46164	**73,44**	0	**0**	45439	**72,29**
YOP57	R1	5621025	0	**0**	1191747	**21,20**	0	**0**	1206262	**21,46**
	R2	5621025	0	**0**	1191747	**21,20**	0	**0**	1195434	**21,27**

Table 2:- Detection of human and parasitic (Pf) reads by PathoExtract and FastQScreen in the unaligned outputs on Plasmodium falciparum during the first experiment.

			PathoExtract				**FastqScreen**			
			Nombre de reads / % de reads							
Échantillon	**R1/R2**	**Total**	**Humain**	**%**	**Pf**	%	**Humain**	**%**	**Pf**	**%**
ANK086	R1	5797080	0	**0**	5796875	**100,00**	0	**0**	5667235	**97,76**
	R2	5797080	0	**0**	5796875	**100,00**	0	**0**	5692058	**98,19**
ANK087	R1	4440445	0	**0**	4440043	**99,99**	0	**0**	4340907	**97,76**
	R2	4440445	0	**0**	4440043	**99,99**	0	**0**	4318264	**97,25**
ANK123	R1	34568	0	**0**	34566	**99,99**	0	**0**	33305	**96,35**
	R2	34568	0	**0**	34566	**99,99**	0	**0**	33148	**95,89**
ANK28	R1	158078	0	**0**	158076	**100,00**	0	**0**	155994	**98,68**
	R2	158078	0	**0**	158076	**100,00**	0	**0**	155361	**98,28**
ANK29	R1	151485	0	**0**	151483	**100,00**	0	**0**	148664	**98,14**
	R2	151485	0	**0**	151483	**100,00**	0	**0**	147951	**97,67**
ANK33	R1	106877	0	**0**	106872	**100,00**	0	**0**	103681	**97,01**
	R2	106877	0	**0**	106872	**100,00**	0	**0**	102593	**95,99**
ANK54	R1	6960	0	**0**	6960	**100,00**	0	**0**	6786	**97,50**
	R2	6960	0	**0**	6960	**100,00**	0	**0**	6746	**96,93**
ANK55	R1	4317	0	**0**	4317	**100,00**	0	**0**	4145	**96,02**
	R2	4317	0	**0**	4317	**100,00**	0	**0**	4107	**95,14**
ANK67	R1	47576	0	**0**	47573	**99,99**	0	**0**	46255	**97,22**
	R2	47576	0	**0**	47573	**99,99**	0	**0**	45876	**96,43**
ANK91	R1	210450	0	**0**	210436	**99,99**	0	**0**	205676	**97,73**
	R2	210450	0	**0**	210436	**99,99**	0	**0**	205385	**97,59**
ANK96	R1	6431181	0	**0**	6431043	**100,00**	0	**0**	6374545	**99,12**
	R2	6431181	0	**0**	6431043	**100,00**	0	**0**	6343999	**98,64**
YOP02	R1	31455	0	**0**	31454	**100,00**	0	**0**	30981	**98,49**
	R2	31455	0	**0**	31454	**100,00**	0	**0**	30828	**98,01**
YOP06	R1	125959	0	**0**	125958	**100,00**	0	**0**	124284	**98,67**
	R2	125959	0	**0**	125958	**100,00**	0	**0**	123696	**98,20**
YOP07	R1	46164	0	**0**	46163	**100,00**	0	**0**	44920	**97,31**
	R2	46164	0	**0**	46163	**100,00**	0	**0**	44543	**96,49**
YOP57	R1	1191747	0	**0**	1191711	**100,00**	0	**0**	1160464	**97,38**
	R2	1191747	0	**0**	1191711	**100,00**	0	**0**	1155330	**96,94**

Proportionate to the number of reads in the "Host unmapped" and "Patho mapped" samples from the first experiment, the pipeline effectively cleaned the data by eliminating human contaminants (0% of reads) while extracting parasitic reads (99% to 100% of reads). The results were validated by comparing the outputs generated by PathoExtract with other tools such as Fastq Screen, thus confirming the accuracy and performance of the pipeline.

Comparison with Other Tools.

We chose to compare PathoExtract, FastQ Screen, HoCoRT, and DeconSeq because, to our knowledge, they represent standards offering comparable methods for contamination detection. Table 3 summarizes the main characteristics of the four software tools compared.

Table 3:- Comparative Table of Features for PathoExtract, FastQ Screen, HoCoRT, and DeconSeq.

	PathoExtract	**Fastq Screen**	**HoCoRT**	**DeconSeq**
Main Objective	Multi-speciesDecontamination	Multi-speciesDecontamination	Human Decontamination	Multi-speciesDecontamination
Types of Contamination	Human, Bacterial, Viral	Human, Bacterial, Viral	Human	Human, Bacterial, Viral
Detection Method	Multi-reference Alignment	Multi-reference Alignment	Human Alignment	Multi-reference Alignment
Fastq/Fastq.gz Input	✔	✔	✔	✔
Paired-end Illumina Support	✔	✖	✔	✔
Command-line Tool	✖	✔	✔	✔
Graphical User Interface (GUI)	✔	✖	✖	✔
Use of Local Resources	✔	✔	✔	✔
Use of Cloud Resources	✔	✔	✔	✔
(Semi-)Automated Installation	✔	✔	✔	✔
Runs on Linux	✔	✔	✔	✔
Quality Control (Trimming/Filtering)	✔	✖	✖	✖
Extraction of Reads Corresponding to Pathogen of Interest	✔	✔	✖	✖
Executable in a single flow	✔	✔	✔	✔
Configurable	✔	✔	✔	✔
Reports	✔	✔	✖	✔
Free for Academic Use	✔	✔	✔	✔

This comparison evaluates four bioinformatics tools: PathoExtract, Fastq Screen, HoCoRT, and DeconSeq based on their main features, strengths, and limitations to determine which is best suited for quality control and multi-species decontamination in studies involving parasites like Plasmodium falciparum. Key evaluation criteria include decontamination capacity, contaminant detection methods, quality control integration, user interface, and pathogen-specific sequence extraction.

PathoExtract stands out as the most comprehensive tool for multi-species decontamination, managing both human and microbial contaminants with precision through multi-reference alignment. Unlike Fastq Screen and HoCoRT, which lack integrated quality control modules, PathoExtract offers a complete solution by eliminating low-quality

sequences prior to decontamination. This feature improves downstream data reliability, making it an all-in-one option, while Fastq Screen and DeconSeq provide multi-species support but without the same level of functionality.

In terms of user accessibility, PathoExtract offers a highly intuitive graphical interface, making it more accessible to users without extensive bioinformatics expertise. This contrasts with Fastq Screen and HoCoRT, which are strictly command-line-based. Additionally, PathoExtract supports pathogen-specific read extraction, streamlining the workflow for pathogen-focused studies, while HoCoRT and DeconSeq lack this capability.

Finally, PathoExtract excels in flexibility and performance, supporting local and cloud-based resources as well as Paired-end Illumina sequences, features not present in all the other tools. Its ability to operate in a single automated workflow and produce detailed reports enhances its efficiency for large-scale NGS analyses. Overall, PathoExtract is the most suitable tool for quality control and decontamination, outperforming Fastq Screen, HoCoRT, and DeconSeq in versatility and performance, making it ideal for complex studies, particularly those involving Plasmodium falciparum.

Discussion:-

PathoExtract is a user-friendly, comprehensive bioinformatics platform designed to streamline the analysis of Next-Generation Sequencing (NGS) data from Plasmodium falciparum-infected samples. Its modular architecture and intuitive graphical interface empower researchers with varying levels of bioinformatics expertise to efficiently manage, process, and analyze complex datasets. By integrating essential features such as quality control, digital filtering, and de novo assembly, PathoExtract ensures accurate and reliable data analysis [13].

A key strength of PathoExtract lies in its ability to effectively filter out host and microbial contaminants, a critical step in studies where sample purity can be compromised [14]. This feature is particularly valuable in clinical and field settings, where obtaining pure parasite isolates can be challenging [15]. Leveraging a sophisticated digital filtering pipeline and robust tools like BWA and Samtools, PathoExtract retains only high-quality, parasite-specific reads. Additionally, the integration of reference genomes further enhances filtering precision, facilitating accurate identification of Plasmodium falciparum sequences.

PathoExtract's de novo assembly module provides a powerful tool for reconstructing parasite genomes from filtered reads [16]. This capability is particularly relevant for research focusing on genomic diversity, drug resistance mutations, and phylogenetic studies. Compared to traditional command-line tools, PathoExtract's streamlined workflow accelerates assembly processes, making it accessible to a broader range of users.

Beyond its core functionalities, PathoExtract offers a user-friendly interface that simplifies the management and analysis of NGS data [17]. The integration of well-established bioinformatics tools into a unified platform empowers researchers with varying levels of expertise. By automating quality control, digital filtering, and de novo assembly, PathoExtract ensures high accuracy and efficiency in handling contaminated samples, where distinguishing between host and parasite sequences is crucial.

While PathoExtract has demonstrated significant benefits, future enhancements could include expanding its compatibility with other parasitic species beyond Plasmodium falciparum. Additionally, incorporating more advanced algorithms (machine learning) could improve data filtering and classification accuracy, especially in cases of mixed infections or low-abundance parasitic reads.

To further broaden its applicability, PathoExtract could benefit from supporting Oxford Nanopore data, which offers long-read sequencing capabilities. This would enable more comprehensive genome assemblies, particularly for resolving complex genomic regions. By expanding its functionality, PathoExtract can address a wider range of research questions and contribute to advancing our understanding of parasitic diseases.

Conclusion:-

This study highlights the utility and versatility of PathoExtract as a comprehensive bioinformatics tool for preprocessing and analyzing NGS data from Plasmodium falciparum-infected samples. Its user-friendly interface allows researchers of all expertise levels to perform essential tasks such as quality control, digital filtering, and de

novo assembly, ensuring the extraction of high-quality parasite-specific data. The tool's automatic removal of host DNA and microbial contaminants is particularly advantageous in clinical and field research contexts.

With the integration of trusted tools like BWA, Samtools, and advanced genome assembly algorithms, PathoExtract provides precise parasite genome reconstruction, making it invaluable for studies on drug resistance, genomic diversity, and phylogenetics. Expanding its compatibility to other parasitic species and supporting long-read sequencing technologies, such as Oxford Nanopore, would further enhance its capabilities. Overall, PathoExtract offers a significant advancement in NGS data analysis, providing a reliable and efficient platform for parasite genomics research.

Acknowledgements:-

We express our deep gratitude for the continuous support provided by the staff of the Parasitology and Mycology Unit and the Genomics and Metagenomics Platform of the Institut Pasteur of Côte d'Ivoire (IPCI) during the testing of PathoExtract. Special thanks go to Dr. Jean René Acquah for his insightful discussions, constructive suggestions, and contributions to improving the double digital subtraction pipeline (Digital Filtering). We also acknowledge Dr. Albert Gnongui for his valuable feedback and comments. Additionally, we extend our sincere appreciation to the National Computing Center of Côte d'Ivoire (CNCCI) for providing the computational infrastructure that facilitated the analysis of NGS sequences in this study through PathoExtract.

We extend our sincere gratitude to the study participants who generously consented to participate in this research. We are also grateful our development partners, particularly Rotary International, for financing the project, and to the entire team of the National Malaria Control Program in Côte d'Ivoire for supplying insecticide-treated nets, antimalarials, and rapid diagnostic tests. We also thank other partners who contributed to the exploration of drug resistance and provided access to technological platforms such as P2M in France and GENEWIZ in the United Kingdom.

Funding:-

This work was funded by Rotary International, whom we thank for their financial support.

Author Contributions:-

SEA designed and developed the application, analyzed the data, prepared the figures and/or tables, drafted or revised the manuscript, and approved the final version. ABA, PNA, PCAA and MBK contributed to the design and development of the application as well as data analysis. JAK and RJ revised the manuscript drafts and approved the final version.

Data Availability:-

PathoExtract is freely available on GitHub at: https://github.com/stanlasso/DREPAL-PATHOEXTRACT.

The data from the DREPAL project are not publicly available. However, they can be provided upon reasonable request and subject to the authors' approval.

The fifteen (15) raw Illumina sequences of public Plasmodium falciparum samples used in this study are accessible via the ENA (European Nucleotide Archive) under the following accession numbers: SRR1654304, SRR1645305, SRR1645306, SRR1645311, SRR1645313, SRR1645314, SRR1645319, SRR1645320, SRR1645321, SRR1645324, SRR1645325, SRR1645327, SRR1645330, SRR1645331, and SRR1645337.

References:-

[1] World Health Organization, « World Malaria Report 2023 », World Health Organization, Geneva, 2023. [Visited on 01-10-2024]. Available at: https://www.who.int/publications/i/item/9789240086173

[2] N. Fukuda et al., « Detection of drug-resistant malaria in resource-limited settings: efficient and high-throughput surveillance of artemisinin and partner drug resistance », Journal of Antimicrobial Chemotherapy, vol. 79, no 6, p. 1418-1422, juin 2024, doi: 10.1093/jac/dkae120.

[3] A. Alcolea-Medina et al., « Unified metagenomic method for rapid detection of microorganisms in clinical samples », Communications Medicine, vol. 4, no 1, p. 135, juill. 2024, doi: 10.1038/s43856-024-00554-3.

[4] P. Ewels, M. Magnusson, S. Lundin, et M. Käller, « MultiQC: summarize analysis results for multiple tools and samples in a single report », Bioinformatics (Oxford, England), vol. 32, no 19, p. 3047-3048, oct. 2016, doi: 10.1093/bioinformatics/btw354.
[5] S. Chen, Y. Zhou, Y. Chen, et J. Gu, « fastp: an ultra-fast all-in-one FASTQ preprocessor », Bioinformatics, vol. 34, no 17, p. i884-i890, sept. 2018, doi: 10.1093/bioinformatics/bty560.
[6] H. Li et R. Durbin, « Fast and accurate short read alignment with Burrows–Wheeler transform », Bioinformatics, vol. 25, no 14, p. 1754-1760, juill. 2009, doi: 10.1093/bioinformatics/btp324.
[7] H. Li et al., « The Sequence Alignment/Map format and SAMtools », Bioinformatics, vol. 25, no 16, p. 2078-2079, août 2009, doi: 10.1093/bioinformatics/btp352.
[8] I. Rumbavicius, T. B. Rounge, et T. Rognes, « HoCoRT: host contamination removal tool », BMC Bioinformatics, vol. 24, no 1, p. 371, oct. 2023, doi: 10.1186/s12859-023-05492-w.
[9] R. Schmieder et R. Edwards, « Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets », PLoS ONE, vol. 6, no 3, p. e17288, mars 2011, doi: 10.1371/journal.pone.0017288.
[10] S. W. Wingett et S. Andrews, « FastQ Screen: A tool for multi-genome mapping and quality control », F1000Research, vol. 7, p. 1338, sept. 2018, doi: 10.12688/f1000research.15931.2.
[11] D. Li, C.-M. Liu, R. Luo, K. Sadakane, et T.-W. Lam, « MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph », Bioinformatics, vol. 31, no 10, p. 1674-1676, mai 2015, doi: 10.1093/bioinformatics/btv033.
[12] J. Köster et S. Rahmann, « Snakemake—a scalable bioinformatics workflow engine », Bioinformatics, vol. 28, no 19, p. 2520-2522, oct. 2012, doi: 10.1093/bioinformatics/bts480.
[13] M. Zanti et al., « Performance evaluation of pipelines for mapping, variant calling and interval padding, for the analysis of NGS germline panels », BMC Bioinformatics, vol. 22, no 1, p. 218, déc. 2021, doi: 10.1186/s12859-021-04144-1.
[14] H. Chen et al., « A comprehensive performance evaluation, comparison, and integration of computational methods for detecting and estimating cross-contamination of human samples in cancer next-generation sequencing analysis », Journal of Biomedical Informatics, vol. 152, p. 104625, avr. 2024, doi: 10.1016/j.jbi.2024.104625.
[15] A. Kilianski et al., « Pathosphere.org: pathogen detection and characterization through a web-based, open source informatics platform », BMC Bioinformatics, vol. 16, no 1, p. 416, déc. 2015, doi: 10.1186/s12859-015-0840-5.
[16] C. Chu, R. Nielsen, et Y. Wu, « REPdenovo: Inferring De Novo Repeat Motifs from Short Sequence Reads », PLOS ONE, vol. 11, no 3, p. e0150719, mars 2016, doi: 10.1371/journal.pone.0150719.
[17] G. Ko et al., « Closha: bioinformatics workflow system for the analysis of massive sequencing data », BMC Bioinformatics, vol. 19, no S1, p. 43, févr. 2018, doi: 10.1186/s12859-018-2019-3.