PROJECT FINAL REPORT

Grant Agreement number: 304875

Project acronym: PATHSEEK

Project title: Automated Next Generation Sequencing for Diagnostic Microbiology

Funding Scheme: Collaborative project*

Date of latest version of Annex I against which the assessment will be made: 8th April 2015

Periodic report: **1st □ 2nd X 3rd □ 4th □**

Period covered: **from 01/03/2014 to 29/02/2016**

Name, title and organisation of the scientific representative of the project's coordinator:

Prof. Judy Breuer, Professor of Virology, Division of Infection and Immunity, University College London

Tel: +44 2031082130

Fax: +44 7053406942

E-mail: j.breuer@ucl.ac.uk

Project website address: www.pathseek.eu

List of Beneficiaries with the corresponding contact names

UNIVERSITY COLLEGE LONDON

Prof. Judith Breuer j.breuer@ucl.ac.uk

ERASMUS UNIVERSITAIR MEDISCH CENTRUM ROTTERDAM

Dr. Martin Schutten m.schutten@erasmusmc.nl

Dr. Saskia L. Smits s.smits@erasmusmc.nl

Prof. Dr. George M.G.M. Verjans g.verjans@erasmusmc.nl

QIAGEN AARHUS AS

Dr. Katja Einer-Jensen Katja.Einer@qiagen.com

OXFORD GENE TECHNOLOGY (OPERATIONS) Ltd

Dr. Graham Speight Graham.Speight@ogt.com

Dr. Jackie Chan Jackie.Chan@ogt.com

Table of Contents

Executive Summary

The overall goal of the PATHSEEK partners (UCL, EMC, OGT and QIAGEN-AAR) was to set up a disruptive diagnostic sequencing platform that would deliver within 24-48 hours clinically relevant information. Towards this aim, the partners have developed a platform comprising of optimised nucleic acid enrichment, next generation sequencing technologies and user-friendly sequence analysis software. This platform can deliver information on all possible drug resistance mutations and supports evaluation of transmission events, directly from a patient sample and within a single assay. We have evaluated this platform using eight key pathogens and reported our findings in the D1.1, D1.2 and D3.1 reports as well as several scientific papers.

The major contribution by OGT was to develop a standard wet-lab protocol which could be applied to all eight test pathogens. Towards this aim, over 1,000 clinical samples were processed at OGT for development work. In addition, alternative library preparation methods were evaluated according to four key performance indicators (KPI's) outlined in the Description of Work. Furthermore, OGT adapted the SOP to enable multiple pathogen detection and co-enrichment of human-biomarkers from a single sample.

The major contribution by UCL and EMC was to establish the beta version of the PATHSEEK platform and compare the results generated with current diagnostic methods. Over 2,000 samples (primarily clinical in origin) were processed at UCL and over 450 samples were processed at EMC. Evaluation of the protocols has shown that whole genome sequences can be obtained at clinically-relevant sensitivities and at read depths that are informative for clinical management and outbreak monitoring for the eight exemplar pathogens.

A general barrier to the diagnostic use of NGS methods has been the challenge posed by the technology and the bioinformatics skills needed to analyse the sequence outputs. The major contribution by QIAGEN-AAR was targeted on the development of a standardised and automated sequence data analysis pipeline with a user interface and reporting format, which addresses both novice as well as experienced bioinformaticians. Significant resources were furthermore spent on development and benchmarking of new bioinformatics tools and analysis workflows using NGS data generated within the project, and key findings have been included in scientific papers.

Although in its current format, the PATHSEEK platform and automated sequence analysis pipelines are not ready for integration into the diagnostic lab, one of the PATHSEEK partners (UCL) intends to establish the sequencing pipeline as a centralised service to academia, diagnostic and public health laboratories. To this extent, the pipeline was successfully used towards the end of the study for the routine diagnosis of antimicrobial resistance and for nosocomial transmission detection in two UK hospitals.

Project context and objectives

Current rapid diagnostic tests for detecting resistance mutations are based on PCR amplification of nucleic acid, followed by Sanger-sequencing. PCR amplifies a single target and for larger pathogens, several reactions are required to cover all the mutations of interest. Other pathogens (for example, Norovirus) are highly variable thus preventing amplification and sequencing of sufficient viral nucleic acid for outbreak analysis unless several reactions are undertaken. Subsequent Sanger-sequencing methods also have their drawbacks as small amounts of sequence data are generated, which further compounds the problems of PCR, especially for detection of drug resistance mutations that are widely spaced or scattered throughout a larger genome. Moreover, Sanger methods are poor for detection of low-level (<5% population frequency) resistance mutations. Finally, in order to generate sufficient sequence data for phylogenetic analysis of outbreaks, Sanger sequencing necessitates too many reactions to be cost and time effective for routine diagnostic use.

Advances in next generation sequencing (NGS) mean that these technologies offer many advantages to diagnostic laboratories as they can carry out massive parallel sequencing thus generating more data. For example, the technology can be harnessed for sequencing multiple different pathogens present in a single sample and/or can be used to simultaneously sequence multiple samples from patients infected with the same pathogen. Incorporating NGS technologies into diagnostic laboratories to sequence whole pathogen genomes has been shown to aid both patient management and public health. The many advantages of sequencing pathogen genomes at high read depths include:

- the concurrent identification of organisms with detection of known and potentially new resistance-conferring mutations,
- detection of mixed infections,
- greater discrimination between strains or genotypes allowing a more accurate determination of transmission events and
- the detection of minor viral populations within an individual.

However, there are major barriers which prevent whole genome sequencing (WGS) being adopted in routine diagnostics, such as the high start-up and running costs and long turn-around time. Nonetheless, as these methods become cheaper and automated, they are starting to have an impact and are being integrated into routine diagnostics and in regional/national public health labs. Leading the way, is sequencing of whole bacterial genomes from cultured isolates which has been used for outbreak monitoring of nosocomial infections with several studies using sequencing to trace transmission routes within hospitals. For real-time outbreak monitoring, this method is most suitable for the fast growing bacterial pathogens such as Staphylococci. However, this approach is not as beneficial for the hard to culture or slow growing pathogens (e.g. Mycobacteria), intracellular bacteria (e.g. Chlamydia) or viruses. The need to culture prior to sequencing means that the same results can be obtained just as quickly from conventional typing methods (e.g. PFGE and MLST). Prior culture may also pose a risk of *in-vitro* selection of specific variants not representing the *ex-vivo* situation potentially leading to misinterpretation of WGS data. The alternative of sequencing directly from clinical samples without prior culture or PCR amplification – known as the metagenomic approach - is too insensitive and does not give good enough data for diagnostic purposes as the vast majority of the sequence reads obtained are of human origin.

To overcome these limitations PATHSEEK, a 42-month project involving four European partners, proposed to set up a disruptive diagnostic technological pathogen sequencing platform that would deliver, in a clinically-relevant turnaround time, all possible drug resistance mutations and nosocomial transmission information, from one patient sample in one single assay. PATHSEEK proposed to develop an end to end assay platform (from sample to analysis) – the PATHSEEK platform - which includes targeted NGS methodology that can be used to sequence whole pathogen genomes directly from clinical samples. To achieve this, small 120bp RNA probes or baits are used to pull out the specific pathogen genetic material, whereas human and commensal DNA is discarded, leaving the pathogen-enriched sample to be sequenced. This method is high-throughput and automatable.

Simultaneous to the development of the “wet-lab” sequencing pipeline, PATHSEEK proposed to develop novel software and bioinformatics tools, SeqFlow, to tackle genuine concerns about data-interpretation complexity and results turn-around times for routine diagnostic applications. This software will facilitate interpretation of results for use by diagnostic lab personnel, clinicians and more experienced bioinformatics users. These measures will provide rapid, personalized results, which are appropriate to real-life diagnostic workflows and clinical needs.

The PATHSEEK platform and the automated sequence analysis pipeline, SeqFlow, would overcome limitations to current rapid diagnostics for detecting resistance mutations which are based on PCR amplification of nucleic acid, prior to Sanger-sequencing. As the entire pathogen genome is sequenced at high read depths, there is simultaneous detection of all known resistance mutations even though these may be widely spaced throughout the pathogen genome. Furthermore, the whole genome sequencing (WGS) data obtained will also facilitate identification of novel resistance-associated mutations in clinical samples of patients refractory to treatment with current antimicrobial drugs. These data can also be used to accurately determine whether isolates from different individuals are related which, in turn can identify nosocomial spread and outbreaks on a national or possibly international scale. Furthermore, if follow-up samples from the same patient are sequenced, the PATHSEEK Seqflow platform will enable monitoring of temporal emergence of resistant variants, thus providing a guided switch to more appropriate antimicrobial drugs.

To evaluate the performance of the proposed platform in clinical diagnostic settings, the consortium has selected eight exemplar pathogens - HIV-1, Hepatitis C (HCV), Hepatitis B (HBV), cytomegalovirus (CMV), norovirus, influenza A (FluA), *Chlamydia trachomatis* (Ct) and *Mycobacterium tuberculosis* (Mtb) - whose characteristics include a clear unmet clinical and public health need and/or for which the threat is of global importance. The selected pathogens, which represent bacteria as well as viruses with DNA and RNA genomes, will be used as proof of concept to show the benefit of new target enrichment on pathogen sequencing and quality.

The PATHSEEK and automated sequence analysis platform could open up a new era in clinical diagnosis and patient management. At the patient level, the rapid delivery of results will ensure early, appropriate therapy thereby helping to reduce the development of drug- resistance and minimizing unnecessary, extra medical expense from the use of inappropriate drugs. Harnessing the platform's multiplex, ultra-deep sequencing capabilities, will also impact public health management of nosocomial infections. At the local hospital level, it will provide, in a simple and efficient manner, all data to control hospital-acquired infections and identify community clusters and transmission chains. At a larger national scale, it will permit European and WHO Reference centers to follow, in real time, the spread of infection, allowing earlier interventions to mitigate pandemics.

To achieve the overall objective, PATHSEEK had the following specific objectives over the full duration of the project:

1. to develop a standardised enrichment and NGS methodology to sequence whole pathogen genomes from clinical samples;
2. to develop software for the analysis of WGS and reporting of clinically relevant data;
3. to test and validate the PATHSEEK platform in clinical diagnostic settings;
4. to assess the potential commercial outputs and establish strategic roadmaps.

Main S&T results/foregrounds

The PATHSEEK workplan was divided into four work packages and the main S&T results are correspondingly broken down by work package for clarity.

Work Package 1: Development of Sequencing Pipeline

WP1 had the following objectives which will be examined in detail below –

1. To develop a benchmark assay: a set of standardised enrichment reagents (Agilent SureSelect) for the capture of whole genomes from eight pathogens (HIV-1, HCV, HBV, CMV, norovirus, FluA, Ct and Mtb) individually.
2. To establish the parameters for standardised sequencing of the above pathogens using Illumina sequencing technology (HiSeq 2000 and MiSeq).
3. To evaluate the performance of this standardised workflow for sequencing of multiple pathogens in the same sample and multiplexed sequencing of many samples in one assay.
4. To evaluate alternative reagents and instrumentation for enrichment, library preparation and sequencing using key performance indicators for comparison with the standardised method above.
5. Using key performance indicators, to choose the most fit-for-purpose assay and platform, with which to develop assays, and robust sequencing workflows for all eight pathogens, for transfer into the UCL research lab as an Alpha-prototype platform at the start of WP3.

Objective 1 – development of a benchmark assay

The eight test pathogens were selected by the PATHSEEK consortium as they all represented a clear unmet clinical and public health need, and the threat posed by them was and is of global importance. These pathogens are a mixture of RNA/DNA viruses and bacteria which vary in genome size from 3 to 4,000 kb (Figure 1) to demonstrate the versatility of the PATHSEEK enrichment technique. During the PATHSEEK project, over 1000 clinical samples were received from UCL and EMC for development work in WP1.

Recovery of whole genome data from a variety of clinical samples

Clinical samples are often a mixture of low amounts of pathogen genetic material and an overwhelming quantity of human and commensal DNA. For example, blood and urine samples contain little-to-no pathogen genetic material whilst stool and saliva are made up of complex mixtures of microorganisms. It has been estimated that there are between 10^{8}-10^{10} bacterial cells per gram of faeces.

We were able to show that it is possible to apply the PATHSEEK sequencing protocol to a range of clinical samples. Using the alpha version of the PATHSEEK sequencing protocol between 500 and 700,000 fold enrichment of pathogen genetic material was achieved. Figure 2 shows the enrichment of CMV DNA from a variety of samples. We found that the ratio of enrichment (% pathogen DNA in enriched sample/% pathogen sequence in original sample) is proportional to the original diagnostic load.

Successful enrichment and sequencing of all eight test pathogens

By M12, we had established a standard method for library preparation incorporating targeted enrichment and sequencing, and applied it to all PATHSEEK pathogens. At least two complete genomes of each pathogen had also been generated as required by milestone 2.

In D1.1 (M18) we reported that we had generated at least eleven complete genomes for each pathogen using the standard method and successfully sequenced *in-vitro*-mixed HIV-1/CMV samples indicating that enrichment of at least two pathogens from a single sample may be possible.

Objective 2 - to establish the parameters for standardised sequencing using Illumina sequencing technology

From the initial work carried out in during months 1-18 we were able to calculate the maximum number of samples in a pool which could be sequenced on an Illumina MiSeq lane (using a V3 300 bp paired-end reagent cartridge). These calculations were based on the experimentally determined percentage of on-target reads (the target being the relevant pathogen) for the two lower genomic input values for which enrichment and sequencing passed QC. It was assumed that a minimum of 20x coverage was required for DNA pathogens, and 800x for RNA viruses.

For the RNA viruses these values ranged from 25 (for norovirus) to 175 (for HCV). As the DNA pathogens have a larger genome size, the number of samples in a pool was lower – between 5 (for CT) and 32 (for CMV). It is worth noting, that these numbers are hypothetical as due to the number of bar-codes available in the Agilent SureSelect^{XT} kit, the maximum number of samples that can be sequenced in a single run is 96.

Objective 3 - to evaluate the performance of this standardised workflow for sequencing of multiple pathogens in the same sample and multiplexed sequencing of many samples in one assay

Optimisation of sample quantification and extraction

Data from the phase I samples enabled us to identify several key issues with the standard workflow as the bacterial pathogens and DNA viruses were generally amenable to the protocol but the RNA viruses were more challenging, typically due to sample degradation. Furthermore, for these samples we found that -

- low amount of sample input RNA in the RNA extraction process reduces RNA yield prior to cDNA synthesis
- measurement of low concentrations of nucleic acid is challenging

To optimise the RNA extraction and cDNA synthesis protocol, we tested four different methods on eight clinical FluA samples (four low and 4 high titre). We found the proportion of target RNA (viral loads) was increased by addition of carrier RNA to the RNA extraction in all samples tested, which resulted in an increase in the proportion of sequence reads mapping to the pathogen of interest. Post ds cDNA synthesis, a column purification step resulted in an increase of the proportion of on-target reads (OTRs) (ie. reads mapping to the pathogen genome) for approximately 30% of the samples. We also assessed cDNA quantification using the Nanodrop and the Qubit in samples extracted with and without carrier RNA. Measurements made with the Nanodrop overestimate the cDNA quantity in samples with carrier RNA, therefore we found that the Qubit method was more accurate/reliable for ds cDNA quantification for these purposes.

From these data we concluded that: (1) RNA samples should be extracted in the presence of carrier RNA, (2) the Qubit ds cDNA assay is required for cDNA quantification, and (3) the purification step post ds cDNA synthesis is not necessary when using the SureSelect^{XT} 3 µg protocol. However, this latter step is essential when using the SureSelect^{XT} 200ng input protocol which was later incorporated as the preferred protocol in the PATHSEEK pipeline.

Additional of bulking agent

At the start of the PATHSEEK project the standard workflow required a DNA input of 3 µg which is not typically obtainable from clinical samples. For samples where DNA/cDNA was <3 µg, addition of human genomic DNA was found to improve sample processing. However, this would have invalidated the co-enrichment of disease-associated host biomarkers, consequently pUC19 was suggested as an alternative bulking agent. We tested a dilution series of cultured FluA bulked with human gDNA and pUC19 and found very little difference between the two bulking agents suggesting that both are suitable for use in the Alpha version protocol.

In early 2014, Agilent released a low input, 200 ng, protocol which meant that not all clinical samples had to be bulked.

Multiple pathogen detection from a single clinical specimen

Initial results from *in-vitro*-mixed HIV-1/CMV samples were promising and suggested that enrichment of two pathogens from a single sample could be possible.

During the course of the project, we were unable to obtain clinical samples from HIV-1 & CMV co-infected individuals as such it was decided to evaluate the co-enrichment of HIV-1 and HCV from eight dual infected patients. We were able to recover a full genome of sufficient coverage in only 1/8 of the samples. To rule out problems with the extraction protocol, we repeated the hybridisation using five laboratory-mixed HIV-1/HCV samples with 1:1 genome copy numbers. We also ran single pathogen controls on the mixed samples to rule out competition between the two pathogen bait sets. As expected, nine of the 10 manually mixed MPD samples passed when hybridised using single pathogen baits. In contrast, only two of the five mixed HIV-1/HCV samples passed for both pathogens; these samples had the highest pathogen input load.

These results suggest that it is possible that the two baits sets are competing against each other during the hybridisation. However, our earlier work with mixed HIV-1/CMV samples show proof of concept.

Human biomarker co-enrichment

A multiplex assay which can co-enrich ds cDNA from HIV-1, FluA or norovirus and the host biomarkers *HLA-B* (HIV-1), *IFITM3* (FluA) and *FUT2* (norovirus) was developed. We tested the combined human marker/pathogen baits on forty-nine clinical samples – 9 norovirus, 24 HIV-1 and 16 FluA, the overall pass rate for samples with successful pathogen and human biomarker enrichment was 57%. All of the FluA samples passed initial QC metrics but only 7/9 (78%) of the norovirus and 5/24 (21%) of the HIV-1 samples generated near-full-length pathogen contigs with >100x depth. Despite the low pass rate for pathogen enrichment, the HIV-1 and FluA samples showed good enrichment of the exons in the human biomarker genes of at least 10x.

We found no enrichment of the *FUT2* gene in three of the norovirus samples and very low depth of coverage in the remaining 6 samples. As there was good coverage of the *FUT2* gene in the other forty clinical samples, the failures are likely to be sample-related instead of due to protocol or bait design. We also found that the coverage profile correlates with known transcripts of the target genes. The intronic portions of the targets showed little-to-no enrichment (no reads mapped to these areas) as would be expected as the baits have captured cDNA not gDNA (which would have been removed during the extraction process). In contrast, the exonic regions showed generally good depth of coverage which ranged from 2 – 2000+ times coverage.

Objective 4 - to evaluate alternative reagent and instrumentation for enrichment, library preparation and sequencing

Testing of alternative library preparation kits

A disadvantage of the PATHSEEK method is the current library preparation turn-around time (TAT) of 3-4 days and minimum starting DNA input of 200 ng and we have evaluated a other library preparation kits in an attempt to reduce overall processing times and/or sample input amounts.

Evaluation of Illumina Nextera XT kit

The Illumina Nextera XT kit offers a one day TAT and has a starting DNA input of 1 ng. We have compared the Nextera XT (Illumina) with the manual Alpha Version PATHSEEK SOP using cDNA/gDNA from cultured HIV-1 and Mtb as exemplar pathogens. Samples with <5 x 10^5 gc/ml for HIV-1 and <5 x 10^4 genome copies Mtb input failed to pass sequencing criteria (<90% reference genome covered) when processed with Nextera XT whilst the PATHSEEK method was able to recover a near-full genome with high depth of coverage. As such we concluded better results were obtained using the PATHSEEK method than for Nextera XT for HIV-1 and Mtb.

Comparison of PATHSEEK SOP with OGT library preparation kit

In August 2015, OGT launched a NGS library preparation kit with a faster protocol and fewer DNA clean-up steps compared to the PATHSEEK SOP. We compared the OGT and PATHSEEK methods using DNA from cultured Mtb and clinical CMV samples and found samples processed with the OGT kit produced similar or better results in terms of % genome recovered (calculated by mapping reads to the reference genome, Mtb H37Rv) and a higher depth of coverage.

Comparison of PATHSEEK SOP with Agilent's Fastprep reagents

The main difference between the PATHSEEK SOP and the Fastprep reagent is the reduction of the hybridization and capture step to 3.5 hours in the fast target system. We carried out a manual library preparation on six samples (three Mtb and three CMV samples) that had been previously sequenced successfully using the standard protocol. We found no noticeable trend between obtaining low or high on-target read percentages from either the PATHSEEK SOP or the fast target system. Overall in five out of the six samples using the SureSelect^{XT} fast target enrichment system did not seem to have a detrimental effect in the amount of usable data that could be retrieved for analysis. However in one CMV sample it did produce a significantly lower read depth across the genome than expected.

Reduced bait set targeting sub-genomic regions associated with antimicrobial resistance in Mtb

To increase the sample throughput of the alpha SOP, we developed a reduced Mtb bait set which would allow more than 16 samples to be sequenced in one MiSeq run (for WGS, between 8-10 Mtb samples can be sequenced in one run). This bait set targets the direct repeat locus (involved in genotyping) and nineteen antibiotic resistance-associated genes. Initial tests using cultured Mtb samples suggested that the reduced bait set successfully enriched the target regions and produces relatively high and even coverage.

We therefore also tested the reduced bait set on clinical sputum samples with previously known fixed drug resistance mutations and two artificial mixtures of three separate Mtb samples with unique drug resistance mutations. We found that the reduced bait set generated high coverage and read depth across target regions with minimal coverage of non-target regions. All previous drug resistance mutations identified by the whole-genome Mtb bait set in the samples tested were also identified at 100% frequency in the reduced bait set results.

Other alternative enrichment methods

The alternative enrichment methods FlexGen and Haloplex PCR are mentioned in the DoW, however, we did not assess these assays. Further investigation into the FlexGen approach found that it was not suitable for PATHSEEK and data from an independent project at OGT found that Haloplex did not perform well with low input samples.

Testing of an alternative sequencing platform for PATHSEEK

Evaluation of the Nanopore MinION was carried out under the Oxford Nanopore Technologies early access program. The Nanopore MinION device uses nanopore-based DNA sequencing with disposable flow cells which are capable of generating reads from long (> 1 kb) DNA fragments. We modified the PATHSEEK SOP to exploit this trait and sequenced cDNA from cultured FluA, gDNA from cultured CMV (amplified by long-range PCR) and two strains of *Mtb* mixed with human gDNA. While unenriched FluA and CMV samples had no reads matching the target organism due to the high background of DNA from the host cell lines, enriched samples had 56.7% and 90.9% on-target reads for the best quality Nanopore reads demonstrating the successful enrichment and sequencing of long DNA fragments.

During PATHSEEK it was not possible to sequence multiple samples by the addition of molecular bar-codes as the platform was and is still in development phase. Due the low capacity and high error rate, particularly in regions of low coverage, we concluded the MinION, at the time of testing (Summer 2015), was not suitable for use in the PATHSEEK workflow.

Objective 5 - to choose the most fit-for-purpose assay and platform for transfer to the UCL research lab as an Alpha-prototype platform at the start of WP3

The most fit-for-purpose assay and platform was outlined in detail the D1.2 – the Agilent Bravo Automated Liquid Handling Platform running a modified version of the SureSelect^{XT} Automated Target Enrichment for Illumina Paired-End Multiplexed Sequencing 200 ng protocol (version F.2) was chosen as the Alpha-prototype platform (Figure 3). This was transferred into the UCL research lab by month 20 for the start of WP3.

In summary, the main achievements of PATHSEEK under WP1 are:

- modified the SureSelect^{XT} protocol to capture pathogen DNA or cDNA from clinical samples
- evaluated alternative library preparation kits and sequencing platforms as and when they became available during the project
- processed over 800 clinical samples
- determined the sensitivity and specificity of the Alpha version PATHSEEK pipeline
- used these data to help train the software developed in WP2

This work has led us to establish a standardised method (SOP) for the enrichment and sequencing of the eight test pathogens both individually and in duplex. The method can also be used to co-enrich pathogens and patient-specific biomarkers, which can be used to alert clinicians to potential adverse effects to particular drug therapies, such as abacavir hypersensitivity.

Deliverables submitted and milestones reached under WP1 are:

D1.1 *Enrichment method for sequencing and detection of multiple pathogens* (submitted M18)
D1.2 *Choice of parameters for Alpha version and standard operating procedures* (submitted M27)

MS1 *Delivery of a definitive panel of target pathogens with verified resistant phenotypes or genotypes* (reached M6)
MS2 *Enrichment and Sequencing of all pathogens* (reached M12)
MS3 *Adequate sensitivity and specificity of Alpha version PATHSEEK prototype* (reached M34)

Work Package 2: Bioinformatics tools and interface

WP2 had the following objectives

1. To develop a new, user-friendly, computationally efficient and accessible software solution for pathogen identification, host biomarker (IL28B&HLAB5701) identification, pathogen variant characterization and molecular epidemiology in a clinical setting.
2. To provide full support for the required data types and workflows in the PATHSEEK solution.
3. To build user interfaces that can be deployed on both stationary and mobile front–end devices such as tablets and phones.
4. To build a flexible computational back-end that can reside on a local installation or on a cloud- or hosted IT-infrastructure.

Objective 1 – To develop a new, user-friendly, computationally efficient and accessible software solution for pathogen identification, host biomarker (IL28B&HLAB5701) identification, pathogen variant characterization and molecular epidemiology in a clinical setting.

Specification of software solution

The requirement specification for the design of the software solution was made by QIAGEN-AAR based on discussions with all partners and included demands on new algorithmic functionality and partner feedback related to user interface and navigation. The computational infrastructure was solely designed by QIAGEN-AAR.

Framework and interface elements

A major barrier to the diagnostic use of NGS methods has been the challenge posed by the technology and the bioinformatic skills needed to analyse the sequence outputs. The major achievement by QIAGEN-AAR was the development of the conceptual framework and interface, which facilitates automated NGS data analysis. To meet the differentiated requirements, the setting provides a condensed summary view targeting the novice users, while a more detailed output including visualization and tables were developed targeting more experienced users. The differentiated reports are kept separately in the web-interface as well as in the generated pdf report.

As more samples were analysed, there was and additional requirement to compare results across different samples. In the final version, the user is able to generate a comparison report which, across the specified samples, aggregates the identified drug resistance, the individual best matching references and the identified key QC numbers.

Some of the above described core elements are now part of the QIAGEN's new GeneReader NGS System, which is the first truly end-to-end solution that brings next-generation sequencing within reach for any lab new sample to insight system.

Development of new algorithms

A core competence of QIAGEN-AAR is development of bioinformatics tools. Within the PATHSEEK project, a new ploidy independent low frequency variant detector optimised towards pathogen populations was implemented in the developed workflows. The identified variants were then compared with an available list of variants known to affect drug resistance.
New epidemiological bioinformatics tools including visualisation of generated phylogenetic trees were also developed to facilitate genotyping. Finally, the ability to provide tack visualization of non-synonymous amino acid mutation was implemented,t hereby enabling genome browsing of detected variants in the context of the mapped read to the reference, as well as detected nucleotide variants.

Host biomarkers

Evaluation of the inclusion of baits against host biomarkers (IL28B and HLA B5701) was performed by OGT using in-house software tools, as the consortium decided during the second period that the automatic analysis workflows should not include characterisation of these host biomarkers.

Documentation (manual for each internal software release) and hands on sessions

In connection to a software release, QIAGEN-AAR has developed/extended the user manual, which was provided simultaneously with installation of the software. Hands-on sessions or Webinars were furthermore provided to introduce new features, improvements and bug fixes. Feedback from partners was collected up to 3 months post installation, and the gathered information was merged with the requirement list and the issues were finally prioritized.

Objective 2 – To provide full support for the required data types and workflows in the PATHSEEK solution.

Proof of concept, a single fixed workflow

The initial proof of concept included a single workflow focused on analysis of Mtb and the analysis workflow included a copy of the publicly available database "Drug Resistance Mutation Database" (http://www.tbdreamdb.com/).

Development of framework for customizing workflows, including resistance and genotyping

As a response to partners' requests, QIAGEN-AAR developed several new tools for the final release that enable experienced users to customize resistance and genotyping parameters according to their needs. The new facilities enabled users to:

a. add a new or modify an existing genotyping sequence list
b. add a new or modify an existing sequence reference list
c. specify custom regions to be included in variant table and extract consensus
d. customize parameter settings within an existing workflow

Framework to build new/supplement existing resistance databases

A lack of publicly available resistance databases was a known limiting factor of the PATHSEEK project. So for the beta release, it was only possible to include a copy of the HIV-1 Drug Resistance Database, Stanford University (http://hivdb.stanford.edu/).

To partners' requests on development of databases related to other pathogens, QIAGEN-AAR developed a framework that enable experienced users to add new or modify drug resistance database as well as genotyping sequence lists. Once pathogen databases are developed and specified in workflows, automatic typing analysis of all eight PATHSEEK pathogens was possible.

Objective 3 – To build user interfaces that can be deployed on both stationary and mobile front–end devices such as tablets and phones.

To ease accessibility and deployment in a clinical setting, QIAGEN-AAR has designed a user-interface that can be offered to the user via web-technologies. The solution works when executed through a standard browser on a computer. However, evaluation of the technology reveals that the web-based solution can be executed and results viewed using portable devices such as smartphones, tablet and notepad computers.

Objective 4 – To build a flexible computational back-end that can reside on a local installation or on a cloud- or hosted IT-infrastructure.

The developed platform and analysis framework enables execution of NGS analysis on a CLC Gx Server through a web based user interface. This construction enables large scale (grid) settings, which supports high throughput sequence analysis.

In summary, the main achievements of PATHSEEK under WP2 are related to one or more of the following categories:

1. Development and validation of new core bioinformatics tools e.g. low frequency variant detector and epidemiological tools
2. Development of a framework that enable experienced users to customizing workflows, including modifying or adding new resistance- and genotyping information.
3. Development of pathogen specific workflows that report resistance- and genotyping of the PATHSEEK pathogens.

Deliverables submitted and milestones reached under WP2 are:

D2.1	*Initial software prototype of core tools and of variant detection tools* (submitted M15)
D2.2	*Alpha software version of full solution excluding epidemiological tools* (submitted M19)
D2.3	*Beta software version of full solutions* (submitted M31)
D2.4	*Release of final software version, including documentation and full manual* (submitted M39)
MS4	*Alpha software version of full solution ready for testing* (reached M19)
MS5	*Beta software version of full solution ready for testing* (reached M32)
MS6	*Complete user friendly software development for diagnostic use* (reached M40)

Work Package 3: Testing and validation of PATHSEEK platform in clinical diagnostic settings

WP3 had the following objectives:

1. To establish the Alpha version of the PATHSEEK in the UCL research laboratory
2. To evaluate key performance indicators for sequencing of the eight pathogens and two host biomarkers on the alpha version of the platform.
3. To establish a Beta version of the PATHSEEK platform in two diagnostic laboratories (EMC and UCL).
4. To evaluate key performance indicators for pathogen sequencing from clinical samples on the beta version PATHSEEK platform

5. To compare key performance indicators for the PATHSEEK platform with current diagnostic methods (e.g. Sanger sequencing for resistance mutations of HIV, HCV, CMV and PCR for Influenza and Mtb)
6. To evaluate performance of the PATHSEEK platform for identifying reinfections, relapses and outbreaks

Objectives 1 and 2– Establish and evaluate alpha version of the PATHSEEK platform in the UCL research laboratory

By May 2014, the alpha version of the PATHSEEK sequencing pipeline had been set up at UCL. At this site, the modified SureSelect^{XT} protocol with 200ng input was used and automated on the Agilent Bravo Automated Liquid Handling Platform based on the requirement for lower amounts of input DNA in the assay, improved consistency and reduced hands-on time.

The criteria agreed by the PATHSEEK consortium for samples to have been successfully sequenced using the alpha PATHSEEK protocol were ≥90% genome coverage and minimum mean read depth of x20 for bacteria and x100 for viruses, with a further requirement for at least two independent reads at a given position. For all pathogens a minimum of 20x depth of coverage was chosen because in our experience this is the depth at which >90% of genome coverage becomes possible, when considering the variability in depth observed over the genome. For small viruses we required an additional depth of at least 100x, to allow calling of minority variants and/or detection of quasi-species which have been well-documented for many RNA viruses, and can have important clinical relevance. A number of samples were sequenced at UCL to evaluate the alpha PATHSEEK protocol and these were included in deliverable 1.2. Applying this methodology, whole genome sequences for eight pathogens from clinical samples were successfully obtained.

Objectives 3 and 4 – Establish and evaluate beta version of PATHSEEK platform at UCL and EMC

Establishment of the beta version

Due to the success of the alpha version of the sequencing pipeline, the beta version comprising the optimal sequencer, reagents and software were installed at EMC (a skilled diagnostic laboratory but with less experience than UCL of using the technology) as well as UCL. This beta version includes:

- Tested extraction protocols for each pathogen;
- Preferred cDNA synthesis protocol for RNA viruses;
- Bait sets designed by the PATHSEEK consortium with continual improvement to capture diversity and complete genome coverage;
- Modified version of the 200ng SureSelect^{XT} protocol – either manual or automated (preferably automated for consistency and high throughout processing of samples);
- Illumina's MiSeq or NextSeq sequencers
- SeqFlow analysis software (the sequencing data generated at EMC was analysed using the PATHSEEK-specific software, SeqFlow. UCL used custom bioinformatics pipelines developed in-house although all versions of SeqFlow-PATHSEEK were tested at UCL and feedback given).

The aim of the PATHSEEK project was to implement a Next Generation Sequencing (NGS) platform based upon bait target enrichment in a diagnostic setting (EMC). Within diagnostic laboratories adherence to a quality system compliant with either clinical diagnostic (ISO15189:2012) or diagnostic testing (ISO17025) procedures is compulsory. The technicians and study supervisor working on the PATHSEEK NGS platform were therefore initially trained in working on Sanger sequencing and NGS sequencing procedures in an ISO15189:2012 compliant laboratory. The study supervisor also obtained further in-depth training from the Dutch board of accreditation into ISO15189:2012 quality systems.

In order to validate the PATHSEEK NGS platform at EMC, it was deemed important to generate data on minority species detection on a pathogen where well-defined materials could be generated for which also a comparator NGS system was available. The initial goal was to use materials similar to clinical materials with respect to nucleic acid composition (chosen from viral particles) and matrix. To this end, FluA H3N2 was chosen. Biological H3N2 clones were generated that differed sufficiently to allow multiple mutation analysis at minority level using triplicate in-vitro culture cloning. In addition, an NGS procedure was set-up on the basis of RT-PCR based target enrichment. However, this validation was abandoned as cultured FluA samples produced poor results compared to clinical FluA samples of similar viral loads. Since the platform aimed at setting-up procedures for clinical samples it was decided -because of the very good results of the platform on clinical samples - not to further optimize for cultured H3N2 samples.

To further validate the PATHSEEK NGS platform it was decided to run a RNA run-off transcript using the PATHSEEK method. A HIV-1 RNA run-off transcript was generated using SP6 as a DNA dependent RNA polymerase. Special attention was paid during the analysis on determining and validating the parameters that need to be set in the analysis pipeline behind the bioinformatics pipeline developed by QIAGEN. After analysis mutants with an average quality below 30 and with a forward/reverse balance below 0.4 were filtered. The analysis showed an exceptionally high number of snip variants (SNP) between 0.1 and 1.0% whichever analysis was used confirming a LLOD for detection of minority variants above 1%.

DNA dependent RNA polymerases are known to dissociate at random from their template and add an extra base (preferably A or T but also G and C) at the end of a transcript. Not surprisingly mainly “A” and “T” insertions and deletions were found; these are not usually present in RT-PCR enriched samples because they require a full-length transcript to be amplified and enriched. Essentially, this confirms the validity of the bait enrichment of the PATHSEEK platform as a relatively unbiased enrichment and detection technique. Furthermore, from these data it can be concluded on the basis of unexpected and unexplained SNP variants detected that preference should be given to (1) removal of duplicate mapped reads, (2) mapping against optimized reference and (3) high insertion/deletion open cost essentially validating the parameters set in the bioinformatics pipeline in the SeqFlow-PATHSEEK Beta version of the software developed within the PATHSEEK project.

Evaluation of the beta version

For the evaluation of the Beta version of the PATHSEEK platform, we used the following criteria to report sensitivity of the assay: for the consensus sequence we continued with the previous criteria used in the evaluation of the alpha version. In addition, we reported the minimum diagnostic value at which 90% of samples give ≥90% coverage at 1x mean read depth to determine the sensitivity of the assay to sequence complete pathogen genomes.

Over 2,000 primarily clinical samples were collected by UCL from numerous diagnostic labs in the UK, Public Health England, and academic collaborators in the UK, South Africa, and Greece for evaluation of the beta version and further optimisations. Over 450 HIV, FluA, HCV, HBV and norovirus samples were sequenced at this site. The beta evaluation was carried out on a wide range of clinical samples including blood and derivatives thereof, urine, stool, cervical/vaginal swabs, sputum, liquid cultures, breast milk, dried blood spots and nasopharangeal aspirates. In addition, for each pathogen a wide range of diagnostic pathogen loads were used in the evaluation: HIV (18 – 4x10^{7} gc/ml), FluA (CT14 - CT 37), norovirus (CT10 – CT43), HBV (110 – 9x10^{8} IU/ml), HCV (12 – 9x10^{7} IU/ml), CMV (50 – 2x10^{7} IU/ml), Mtb (2x10^{3} – 9.6x10^{6} gc/ml), Ct (1x10^{4} - 8x10^{8} gc/ml).

As expected, there was a positive correlation between diagnostic pathogen loads and the percentage on-target (ie. reads mapping to the reference genome) for all pathogens. The current lower limits of the assay for sequencing the complete genome of each pathogen (i.e. minimum diagnostic value at which 90% of samples give ≥90% coverage at 1x, 20x (for bacteria and CMV) or 100x (for small viruses) mean read depth (RD)) are shown in Table 1.

For the small viruses (with the exception of FluA), there is a significant difference in the sensitivity of the PATHSEEK platform for samples sequenced at UCL and EMC, with higher sensitivities achieved at UCL. Possible reasons for this include (i) automated versus manual extraction protocols employed by UCL and EMC respectively, (ii) different bioinformatics workflows used (UCL used custom designed pipelines whereas EMC used SeqFlow-PATHSEEK which is primarily designed for drug resistance analysis and not optimised for sensitivity analysis); (iii) more samples have been sequenced at UCL (a higher proportion of which have high pathogen loads which may skew the sensitivity results) and, finally; (iii) UCL has more experience of the PATHSEEK platform.

In order to report on the sensitivity of the assay for reporting variants in target regions (ie. regions of the pathogen genome known to be associated with antibiotic and/or resistance mutations), we initially carried out a theoretical analysis to determine the parameter thresholds required to minimise the false positive variant calling rate whilst remaining as sensitive (minimising false negatives) as possible. This is highly dependent on the rate of error (sources of which are PCR and sequencing errors) in our system and the depth of coverage. It was concluded that a minimum of 5 reads and 2% frequency would be the most conservative thresholds that could be applied across a range of sequencing depths. However depending on the specific analyses (whether diagnostic in a clinical setting or academic), lower thresholds such as 1% frequency could also be applied at depths greater than 1000x. Thus, for the evaluation of the Beta version of the PATHSEEK platform, minimum diagnostic values at which 90% and 50% of samples have 100% coverage at read depths of x25 (allowing calling of SNPs at 20%) and x250 (allowing calling of SNPs at 2%) were reported.

Target regions sensitivities have been determined for a subset of viral samples as shown in Table 2.

Objectives 5 and 6 – Comparison of the PATHSEEK platform with current diagnostic methods and evaluation of the performance of the platform

Comparison with current diagnostic methods

Apart from Ct, for which there is currently no molecular test for the detection of resistance, various molecular assays are used by diagnostic labs to detect and identify pathogens and to identify mutations conferring resistance from clinical samples. These include real-time PCR for the detection of viruses, PCR followed by Sanger sequencing for identification of resistance mutations in viruses, molecular tests (such as Cepheid's GeneXpert MTB/RIF) and nucleic acid amplification assays for the detection of chlamydia. The disadvantages of these assays include (i) information provided is limited to particular target regions and not all resistance-conferring mutations are detected, (ii) to capture genotype diversity and widely-spaced regions conferring resistance, multiple PCR reactions of required, (iii) low level minority variants occurring at <20% frequencies that may be relevant for treatment response, may not be detected and, finally (iv) some assays do not provide phylogenetic analysis to identify outbreaks and epidemics.

In comparison, the advantages of using high throughput NGS technologies to sequence whole pathogen genomes in a single reaction, is the simultaneous identification of organisms, detection of known and potentially new resistance-conferring mutations, detection of mixed infections, and greater discrimination between strains or genotypes allowing a more accurate determination of transmission events and the detection of minor viral populations within an individual. NGS technologies, particularly for sequencing bacterial genomes from cultured isolates, are beginning to become integrated into routine diagnostics and in public health labs. In addition, alternative NGS methods such as

(i) a target enrichment approach in which pathogen nucleic acid is partially purified using pathogen-specific baits prior to sequencing, ie PATHSEEK method;

(ii) a 'metagenomics' approach in which the entire sample is sequenced and the human DNA reads are discarded during the subsequent analysis; and

(iii) a PCR approach in which the entire pathogen genome is amplified on overlapping PCR fragments, which are then pooled and sequenced;

are being developed which remove the need for culture which are beneficial for the hard to culture or slow growing pathogens, intracellular bacteria or viruses. An advantage to the metagenomics approach is that the data obtained is unbiased by any processing steps although the majority of the reads will be of human origin. The PCR approach is relevant for small viral genomes but would be impractical for larger viral or bacterial genomes. During PATHSEEK we have undertaken various comparisons of these approaches with collaborators and have shown that the PATHSEEK method is less prone to failure, particularly for diverse and possibly degraded samples, than the other approaches. These comparison studies include:

- comparing the three approaches for WGS of HCV directly from clinical samples;

- generating >90% genome coverage at >100x average read depth for 52 out of 80 HIV-1 samples from the 1980s. Seven of these samples sequenced using the overlapping PCR based approach achieved 3 – 80% genome coverage whilst sequencing of the same samples using the PATHSEEK approach achieved 72 – 100% genome coverage.
- comparing the targeted enrichment approach versus the PCR/sequencing approach for genotyping norovirus.
- enrichment of ancient Mtb from mummified bone samples achieved better genome coverage than the metagenomics method.

PATHSEEK has shown that complete genomes can be obtained even from low titre samples and at read depths to be informative both for clinical management and public health strategies from all eight exemplar pathogens. However, despite these achievements, the following barriers to the PATHSEEK platform being successfully incorporated into diagnostic labs for clinical diagnostics remain:

- High set up and running costs.
- Well-trained personnel are required to generate good quality sequencing data and achieve clinically-relevant sensitivity of the platform.
- The turnaround time from sample receipt to sequencing analysis result is currently too long to influence patient management.
- Curated databases of multiple genotypes and resistance-associated mutations for each pathogen are required for analysis.

Identifying relapses, reinfections and outbreaks

We have used the PATHSEEK sequencing pipeline for routine diagnosis of antimicrobial resistance in Mtb, CMV and for nosocomial transmission detection for norovirus and FluA from outbreaks at Great Ormond Street (GOSH) and Norfolk and Norwich University Hospitals (NNUH). This work has informed both treatment regimens, with the WGS data confirming the suspicions of clinicians that a patient had CMV resistant to Ganciclovir, and infection control, indicating that suspected norovirus transmission between patients at Great Ormond Street was not in fact occurring. For FluA samples tested from NNUH we could identify transmission between patients based on phylogenetic relatedness. Similar analysis was carried out for the FluA samples from GOSH and along with possible transmission events we could also identify the rise and disappearance of drug resistance in patients that correlated well with dates of drug administration.

In summary, the main achievements of PATHSEEK under WP3 are:

- Establishing the Beta version of the PATHSEEK platform at UCL and EMC
- Evaluating the Beta version of the PATHSEEK platform including high throughput sequencing of the eight PATHSEEK pathogens from clinical samples

- Comparison of the PATHSEEK platform with current diagnostic methods.
- Evaluation of the performance of the PATHSEEK platform for identifying outbreaks

Deliverables submitted under WP3 are:

D3.1 *Completion of performance standards for the Beta version* (submitted M39)
D3.2 *Final report on the performance of the PATHSEEK platform* (submitted M42)

Milestone MS7 *Achieve 24-48hr turnaround enrichment and sequencing time* was not reached. This was an ambitious aim and the current turnaround time is approximately 5-7 days. This is partly due to processing delays caused by sample batching. We have looked at alternative methods to reduce current processing times particularly by introducing shorter bait hybridisation times. The two methods with the most successful outcomes (OGT library prep kit and Agilent's Fastprep reagent kit) are currently only in manual format so are unable to be incorporated into the automated platform.

Work Package 4: Innovation, exploitation and dissemination

WP4 had the following objectives:
1. IP management and protection
2. To establish strategic roadmaps
3. Dissemination and promotion of the results

Objective 1 – IP management and protection

The Intellectual Property Committee for the PATHSEEK project, which has representation from each partner, have monitored the progress of the project as part of the regular consortium meetings, in order to capture any outputs that should be patented or protected by other means. As the project has progressed, the group have decided that there were no direct outputs conducive to patent protection, especially when considering the existing prior-art within the field. Without doubt the "know-how" developed by the consortium as a whole is of value and can be used for future endeavours.

Objective 2 – Establish strategic roadmaps

An intellectual and industrial knowledge management report was submitted early on in PATHSEEK followed by a strategic roadmap which was initially developed to outline the key waypoints from the current PATHSEEK research through to product launch. However, in 2014 OGT concluded that the commercialisation of PATHSEEK in a kit/service form was not relevant to them at this time and similarly QIAGEN decided that they were not interested in commercialising the PATHSEEK platform. The commercialisation focus therefore shifted and a second strategic roadmap was developed by UCL.

There is scope to develop PATHSEEK as a research, laboratory or diagnostic tool in a number of areas in the healthcare and life sciences field, and a detailed assessment of each proposed commercial/clinical offering was carried out. In summary, there is an immediate opportunity to offer the PATHSEEK enrichment and sequencing process as a service in the research community. UCL have already begun exploring this option as they have established the workflow and have a suitably skilled team, who will be able to deliver this service. Whilst the market size is unknown, there is little commercial opportunity for the other partners of this project; thus it is anticipated they will not continue to support this effort.

As WGS technologies become less expensive and faster, and the methods more straight-forward, we believe that pathogen WGS directly from clinical samples (ie. without prior culture) will have application in diagnosis. Adoption into clinical or reference laboratories is a medium term aim. This will require some strategic partnerships and collaborations to improve and validate the PATHSEEK platform and its workflow before integration. We have confidence that this can be achieved given the ease at which the beta version was established at EMC as part of this project. However, the lower sensitivities achieved at EMC demonstrate that personnel training and expertise is essential. Revenues needed to validate the PATHSEEK platform are likely only to be generated after further development and clinical research endeavours.

Finally, the long term vision proposes that a diagnostic device could be developed to enable real-time pathogen WGS. There are a number of companies and research groups who are exploring and working on technological solutions to provide innovative sequencing platforms in order to expedite time to result. The most well-known of these is the MinION produced by Oxford Nanopore – we have tested this as part of this project as it is reportedly good for long reads, but in its current form it was not found to be sensitive or accurate enough for this workflow. Other companies have developed devices that read short targeted regions to act as a diagnostic to confirm a specific infection or disease; these however are not able to read whole genomes so for the PATHSEEK workflow these are not applicable. It is promising to see that such advances have already been made and looking ahead the ideal would be to reach this goal. Given the current developments, it is likely to take at least 10 years before any potential revenue stream could be developed. The only way to take this idea forwards would be to leverage further research funding and develop collaborations with knowledgeable partners to explore the possibilities.

Objective 3 – Dissemination and promotion of the results

To date the project has disseminated results via peer review publications, oral and poster presentations given at meetings and conferences, short articles and press releases.

A lot of interest was generated from our paper detailing whole genome sequencing of Mtb directly from sputum. During WP1, we were able to produce sequence data from twenty-four smear-positive sputum samples, of which 20 were high quality (> 20x depth and >90% of genome covered). We found a high level of concordance between the phenotypical resistance and predicted resistance based on genotype. These data were published in May 2015 in the Journal of Infectious Diseases (D.O.I 10.1128/JCM.00486-15).

This was followed up by a review of the potential for WGS of multi-drug resistant TB in Expert Review of Anti-Infectious Therapy in 2015 (D.O.I 10.1586/14787210.2016.1116385).

A similar paper by PATHSEEK detailing whole genome sequencing of Ct directly from urine and vaginal swab samples was published in 2014. In this paper we showed that the PATHSEEK method enabled sequencing of full *Chlamydia* genomes from clinical specimens at >10-fold sensitivity than had previously been reported. To evaluate if targeted enrichment introduces any mutational bias, we compared the single nucleotide polymorphic differences (SNPs – single nucleotide polymorphisms) found between the consensus sequence recovered from a sample processed with enrichment and the consensus sequence recovered from the same sample processed without enrichment. Overall, no differences in the SNP profiles were found at any position in the genome (both coding and noncoding position) indicating that no mutational bias is introduced.

In addition, at least 11 further papers, commissioned reviews or book chapters arising from work during PATHSEEK are in preparation or have already been submitted.

The website has also been kept up to date with news and events at which PATHSEEK members were attending. This already has raised awareness of the achievements to date and as such the consortium has already been approached by a number of interested parties, both academic and commercial enquiring as to the availability and use of PATHSEEK going forwards.

The consortium has produced marketing material in collaboration with Research Media (RM) for the general public. This article was published on 17th December and disseminated broadly by RM to over 100,000 global contacts; from researchers, academics and universities to funding agencies, policy makers, NGOs and the general public.

Deliverables submitted and milestones reached under WP4 are:

D4.1 *Intellectual and industrial knowledge management report* (submitted M6)
D4.2 *Report on strategic roadmaps and marketing materials* (submitted M40)
MS8 *Strategic roadmap* (reached M18)

Potential impact

Infection remains among the ten leading causes of mortality in high-income countries, but particularly in low- and middle-income countries. While mortality from infection is falling in Europe, morbidity remains high. Malaria, TB, HIV and respiratory infections are still the major causes of mortality worldwide. In addition, new infection-related problems are emerging both globally and in Europe. The increase in resistance to antibiotics and antivirals, infections resulting from complex modern medical interventions and the impact of burgeoning new health technologies in the fields of cancer, transplantation and others, on infection, present ever greater challenges for infectious diseases in the 21st century.

With the advent of NGS technologies, exciting new opportunities for the use of whole genome pathogen sequencing in diagnostic microbiology and virology have arisen. These technologies allow us to sequence whole pathogen genomes in a single reaction, enabling the simultaneous identification of organisms, detection of known and potentially new resistance-conferring mutations, detection of mixed infections, and greater discrimination between strains or genotypes allowing a more accurate determination of transmission events and the detection of minor viral populations within an individual. The information generated from WGS can be used to inform treatments given to patients (e.g. switch to more appropriate antimicrobial therapy due to presence and/or emergence of resistant variants) and also for infection control (e.g. outbreak management).

Major barriers to WGS being adopted in routine diagnostics are high cost and long turn-around times. However, as these methods become cheaper and automated, they are beginning to have an impact and become integrated into routine diagnostics and in public health labs. Leading the way, is sequencing of whole bacterial genomes from cultured isolates which has been used for outbreak monitoring of nosocomial infections with several studies using sequencing to trace transmission routes within hospitals.

However, this approach is not as beneficial for the hard to culture or slow growing pathogens (e.g. Mtb), intracellular bacteria (e.g. Chlamydia) or viruses. The need to culture prior to sequencing means that the same results can be obtained just as quickly from conventional culture methods. This may also pose a risk of *in-vitro* selection of specific variants not representing the *ex-vivo* situation potentially leading to misinterpretation of WGS data. PATHSEEK was initially conceived to overcome this problem by generating and analysing deep sequencing data directly from a clinical specimen.

During PATHSEEK, we have optimised nucleic acid enrichment and next generation sequencing technologies and developed user-friendly sequence analysis software for use in the diagnostic environment that would support interpretation, by diagnostic scientists and clinicians, of the sequence data and provide clinical reports. PATHSEEK has shown that whole genome sequences can be obtained at clinically-relevant sensitivities and at read depths that are informative for clinical management and outbreak monitoring for the 8 exemplar pathogens. Although in it’s current format, the PATHSEEK platform is not ready for integration into the diagnostic lab, one of the PATHSEEK partners (UCL) intends to establish the sequencing pipeline as a centralised service to academia, diagnostic and public health laboratories. To this extent, the pipeline was successfully used towards the end of the study for the routine diagnosis of antimicrobial resistance and for nosocomial transmission detection in two UK hospitals.

The capability to detect drug resistance within a single assay would allow prompt treatment with the appropriate drug(s). This will improve clinical outcomes, reduce adverse drug reactions, and enable prompt detection of resistance. Detection of emerging resistance through monitoring of low level resistance mutations would also inform clinical treatment. PATHSEEK has shown that mixed infections can be detected from a single sample, making possible the merging of microbiology resistance testing onto one platform. This could impact not only on patient management, especially where bacterial and viral infections occur together in the same patient and even the same sample, but on diagnostic costs and workflow. While diagnostic work streams for virology and microbiology are currently separated, the PATHSEEK platform will have the capacity and versatility to allow unified simultaneous processing of multiple pathogens and samples, further adding to its disruptive potential.

The second and equally important health impact from PATHSEEK will be the generation of pathogen genetic data for surveillance of transmission contacts (Ct, Mtb, and HBV), nosocomial transmission (Norovirus) and outbreaks (FluA, Mtb, and HBV). Rapid identification of nosocomial transmissions will inform the timing and nature of infection control measures and impact on control of hospital acquired infections.

We recently undertook a comprehensive evaluation of the potential of the PATHSEEK platform in the diagnostic or reference lab (D4.2). We concluded, that as WGS technologies becomes less expensive and faster, and the methods more reliable and straight-forward we foresee that pathogen WGS will have application in this area, be adopted as routine method and in turn may guide clinical practice. In its current formulation, the PATHSEEK platform could be used as a service facility where samples were referred to a centralised lab.   Adoption of the PATHSEEK platform into clinical or reference laboratories is a medium term aim, with a long term vision of a diagnostic device could be developed to enable real-time pathogen WGS.

Tables & Figures

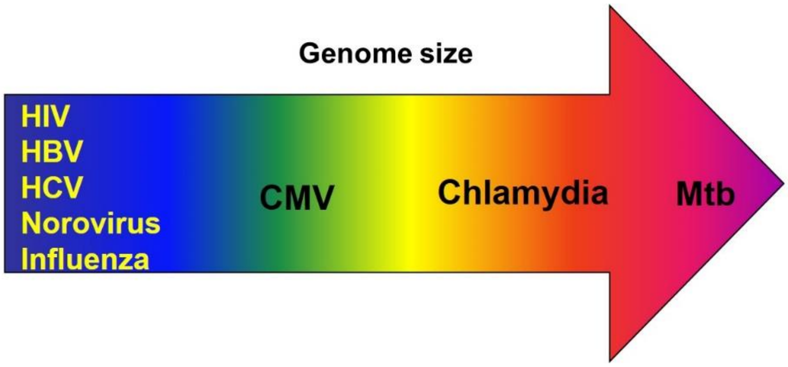

Figure 1. A spectrum of bacterial and viral pathogens of significant clinical relevance, were chosen to assess the robustness of the technology from small RNA viruses to large bacterial DNA genomes.

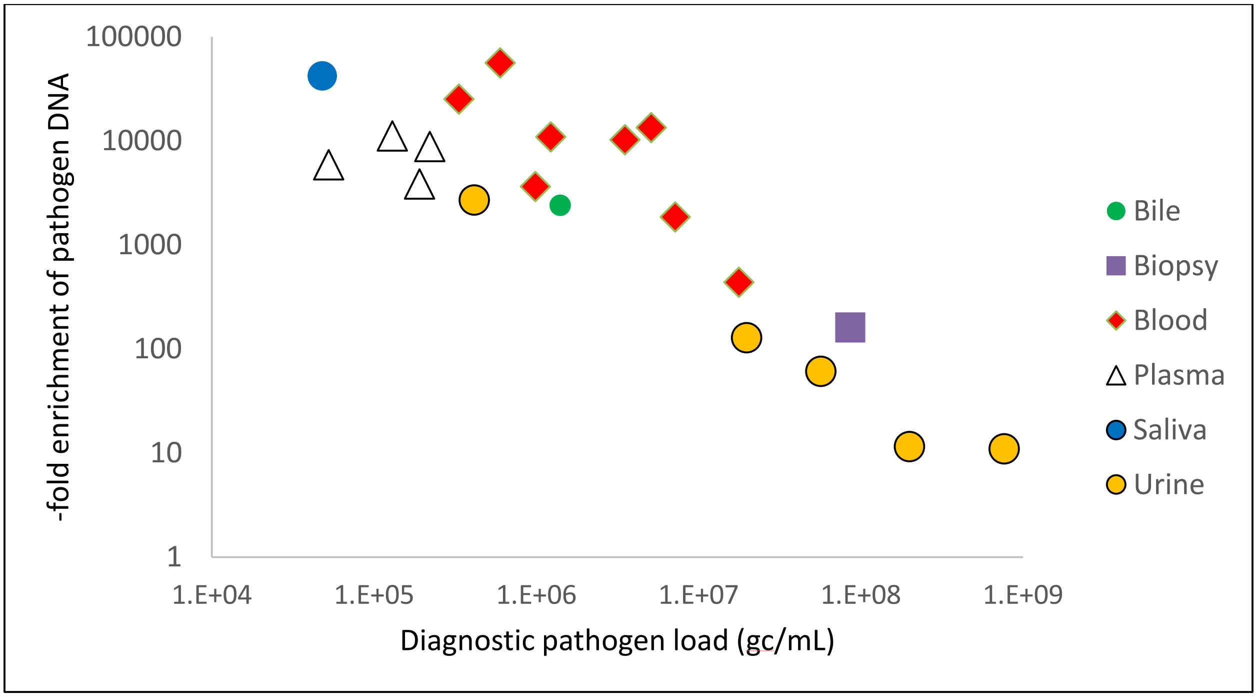

Figure 2. The relationship between diagnostic pathogen load and the ratio of enrichment from 20 CMV samples.

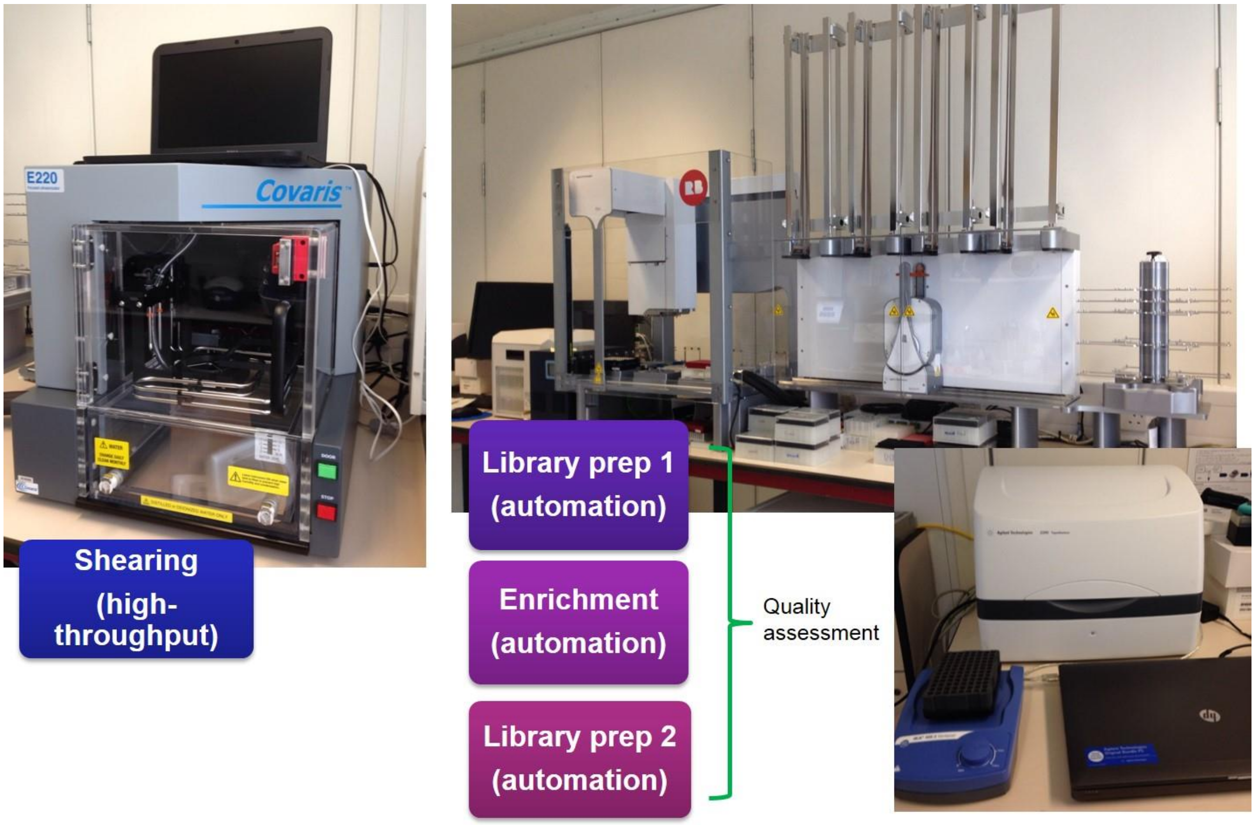

Figure 3. The Alpha-prototype platform in the UCL research lab.

Pathogen	Evaluation at UCL				Evaluation at EMC			
	Number of samples	x1	x20	x100	Number of samples	x1	x20	x100
Chlamydia trachomatis	16	$3.51x10^5$ gc/ml	$5.34x10^5$ gc/ml	-	-	-	-	-
	24	Ct 30		-				
Mycobacterium tuberculosis	13 (London)	1610 gc input *	1610 gc input *	-	-	-	-	-
	18 (SA)	$1.87x10^5$ gc input *	$1.87x10^5$ gc input *	-				
CMV	111	3.35 $x10^4$ gc/ml	5.04 $x10^4$ gc/ml	-	-	-	-	-
HBV	120	3 x 10^2 IU/ml	-	3.2 x 10^2 IU/ml	81	$1.4x10^8$ IU/ml	-	$1.4x10^8$ IU/ml
HCV	352	12 IU/ml	-	12 IU/ml	70	$6.23x10^6$ gec/ml	-	$6.23x10^6$ gec/ml
HIV	71	89 IU/ml	-	89 IU copies/ml	46	7.9 x 10 gec/ml	-	1.02 x 10^5 gec/ml
Influenza A	111	Ct 37	-	Ct 34	30 H1N1	Ct 33	-	Ct 33
					96 H3N2	Ct 40	-	Ct 26.5
Norovirus	563	Ct 43	-	Ct 43	146	Ct 24.5	-	NA**

Table 1. Sensitivity of the PATHSEEK Beta version for sequencing whole genomes of each pathogen directly from clinical material. The minimum diagnostic values at which ≥90% of samples have ≥90% genome coverage at 1x, and either x20 (bacteria/CMV) or x100 (small viruses) mean read depths are shown. * for the Mtb samples, the minimum number of gc copies input into library preparation is shown; ** , Norovirus samples of high titer and specific genotypes (gII.14, gII.17 & gI.6) precluded sensitivity calculation due to low coverage of reference sequences provided in the database of the SeqFlow beta software

Pathogen	Evaluation at UCL				Evaluation at EMC			
	90% samples		50% samples		90% samples		50% samples	
	x25	x250	x25	x250	x25	x250	x25	x250
CMV	1.19 x 10^6 gc/ml	nd^{*1}	1.31 x 10^5 gc/ml	nd^{*1}	-	-	-	-
HBV	4.2 x 10^4 IU/ml	1.2 x 10^5 IU/ml	1.4 x 10^3 IU/ml	8.3 x 10^3 IU/ml	-	-	-	-
HCV	36 IU/ml	36 IU/ml	36 IU/ml	36 IU/ml	-	-	-	-
HIV	920 copies/ml	2 × 10^4 copies/ml	920 copies/ml	920 copies/ml	nd^{*2}			
Influenza A	Ct 18	Ct 18	Ct 34	Ct 34	H1N1: Ct33 H3N2: Ct22.2	H1N1: Ct33 H3N2: Ct22.2	H1N1: Ct33 H3N2: Ct34.4	H1N1: Ct33 H3N2: Ct34.4

Table 2. Sensitivity of the PATHSEEK Beta version for sequencing of the viral target regions directly from clinical material. The minimum diagnostic values at which either ≥90% or ≥50% of samples have 100% coverage at x25 (variant calling at 20%) and x250 (variant calling at 2%) read depths are shown. Norovirus is not included as there are no specified target regions for this pathogen. nd^{*1}, none of the samples tested reached this criteria; nd^{*2}, target regions not analyse