University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Low, SJ;O'Neill, M;Kerry, WJ;Wild, N;Krysiak, M;Nong, Y;Azzato, F;Hor, E;Williams, L;Taiaroa, G;Steinig, E;Pasricha, S;Williamson, DA

Title:
PathoGD: an integrative genomics approach to primer and guide RNA design for CRISPR-based diagnostics

Date:
2025-12-01

Citation:
Low, S. J., O'Neill, M., Kerry, W. J., Wild, N., Krysiak, M., Nong, Y., Azzato, F., Hor, E., Williams, L., Taiaroa, G., Steinig, E., Pasricha, S. & Williamson, D. A. (2025). PathoGD: an integrative genomics approach to primer and guide RNA design for CRISPR-based diagnostics. Communications Biology, 8 (1), pp.147-. https://doi.org/10.1038/s42003-025-07591-1.

Persistent Link:
https://hdl.handle.net/11343/359764

License:
CC BY-NC-ND

communications biology

Article

https://doi.org/10.1038/s42003-025-07591-1

PathoGD: an integrative genomics approach to primer and guide RNA design for CRISPR-based diagnostics

Check for updates

Soo Jen Low [1,2] ✉, Matthew O'Neill [1,3], William J. Kerry [3], Natasha Wild [3], Marcelina Krysiak[1], Yi Nong [1,2], Francesca Azzato[1,2], Eileen Hor[2], Lewis Williams [3], George Taiaroa[1], Eike Steinig[1,2], Shivani Pasricha [1,3,6] ✉ & Deborah A. Williamson [1,4,5,6]

Critical to the success of CRISPR-based diagnostic assays is the selection of a diagnostic target highly specific to the organism of interest, a process often requiring iterative cycles of manual selection, optimisation, and redesign. Here we present PathoGD, a bioinformatic pipeline for rapid and high-throughput design of RPA primers and gRNAs for CRISPR-Cas12a-based pathogen detection. PathoGD is fully automated, leverages publicly available sequences and is scalable to large datasets, allowing rapid continuous monitoring and validation of primer/gRNA sets to ensure ongoing assay relevance. We designed primers and gRNAs for five clinically relevant bacterial pathogens, and experimentally validated a subset of the designs for detecting *Streptococcus pyogenes* and/or *Neisseria gonorrhoeae* in assays with and without pre-amplification. We demonstrated high specificity of primers and gRNAs designed, with minimal off-target signal observed for all combinations. We anticipate PathoGD will be an important resource for assay design for current and emerging pathogens. PathoGD is available on GitHub at https://github.com/sjlow23/pathogd.

Rapid and accurate diagnosis is essential for the prevention and control of infectious diseases. However, traditional methods such as polymerase chain reaction (PCR) are generally restricted to the laboratory setting due to their reliance on trained personnel and expensive equipment[1]. Accordingly, there is a critical need for affordable point-of-care (POC) diagnostics to increase access to and adoption of testing[2]. One approach that has gained momentum in recent years leverages clustered regularly interspaced short palindromic repeats (CRISPR) technology. CRISPR technology has found significant applications in infectious disease diagnostics owing to its simplicity, high specificity, and non-reliance on sophisticated equipment[3]. CRISPR-based diagnostics have been used for the detection of viruses, bacteria, and parasites, including Zika and Dengue virus[4,5], human papillomavirus[6], *Mycobacterium tuberculosis*[7], *Plasmodium falciparum*[8,9] and more recently, SARS-CoV-2[10–12] and *Monkeypox virus*[13,14]. To achieve pathogen detection at low concentrations, CRISPR is often coupled with an upstream isothermal amplification step such as recombinase polymerase amplification (RPA) or loop-mediated isothermal amplification (LAMP).

CRISPR-based diagnostic assays involve a multi-stage process that exploits the unique properties of Cas12 and Cas13 nucleases. Cas nucleases are directed to bind their target regions by a guide RNA (gRNA) specifically designed to complement the target sequence of interest. In addition to cleaving its target DNA (Cas12) or RNA (Cas13), the Cas nucleases also non-specifically cleave single-stranded DNA (ssDNA) and single-stranded RNA (ssRNA) through target-activated collateral cleavage activity[6,15–17]. Labeling ssDNA and ssRNA molecules with fluorophore-quencher pairs allows the detection of target nucleic acids through the fluorescent signal generated by the cleavage of reporter molecules. The versatility of the CRISPR-Cas system lies in its ability to be tailored for the detection of nucleic acids from virtually any pathogen through the design of specific gRNAs that complement the target region.

Crucial to the success of CRISPR-based diagnostics is the selection of a diagnostic target site, followed by the design of primers and gRNAs highly specific to the target. Identification of reliable targets and subsequent assay design can be labor-intensive and typically involves careful analysis of genomic sequences, assessment of conserved regions, consideration of off-

[1]Department of Infectious Diseases, The University of Melbourne at The Peter Doherty Institute for Infection and Immunity, Melbourne, VIC, Australia. [2]Victorian Infectious Diseases Reference Laboratory, The Royal Melbourne Hospital at The Peter Doherty Institute for Infection and Immunity, Melbourne, VIC, Australia. [3]Walter and Eliza Hall Institute of Medical Research, Parkville, Melbourne, Australia. [4]School of Medicine, University of St Andrews, Fife, Scotland. [5]MRC-University of Glasgow Centre for Virus Research, Glasgow, Scotland. [6]These authors contributed equally: Shivani Pasricha, Deborah A. Williamson.
✉e-mail: soojen.low@unimelb.edu.au; shivani.pasricha@unimelb.edu.au

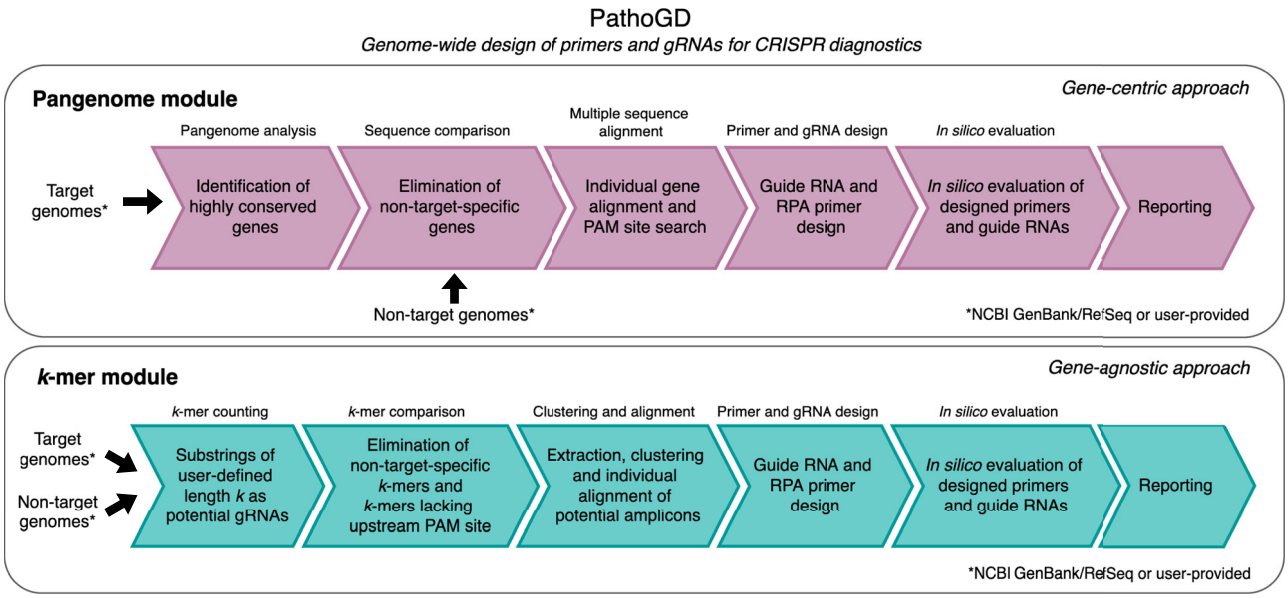


Fig. 1 | Overview of the PathoGD pipeline. The main steps involved in the pangenome and *k*-mer modules are highlighted. Further details about each step are illustrated in Supplementary Fig. 1.

target effects, and iterative refinement to achieve optimal performance. Consequently, conventional diagnostic assays often rely on a limited set of marker genes as target regions for amplification, established prior to the next-generation sequencing era, such as *porA* in *Neisseria gonorrhoeae*[18] and *nuc* in *Staphylococcus aureus*[19]. This standard approach may fail to capture the full genetic diversity of a pathogen, potentially resulting in a diagnostic escape that is only detectable when using a different diagnostic platform[20]. CRISPR–Cas12a-based assays also require the presence of a protospacer adjacent motif (PAM), TTTN directly upstream of the gRNA sequence for recognition and DNA binding[21], potentially limiting the use of some well-established diagnostic markers.

Several tools for gRNA design have been developed and are available through a web interface or as a command-line tool. However, most of these were tailored for genome editing applications, primarily in eukaryotic organisms using Cas9. Additionally, some are restricted to a single genome or certain model organisms or contain pre-defined databases, thus limiting their utility, especially for bacterial pathogens. These tools include but are not limited to, DeepCRISPR[22], GuideMaker[23], CRISPRscan[24], CRISPR-DT[25], CRISPRseek[26], GuideScan[27], and crisprVerse[28]. The gRNA design principles for genome editing applications differ from those for CRISPR-based nucleic acid detection or diagnostic applications in multiple ways. In genome editing, the gRNA should bind to a single target site in the genome, whereas multiple gRNA target binding sites in a genome are often advantageous for diagnostic applications. Off-target effects in genome editing are evaluated within the same genome to predict mutations at unintended genomic loci, whereas genomes from different species are often used for evaluating gRNA specificity in diagnostics. Tools developed specifically for the design of gRNAs in diagnostics are PrimedSherlock[29], CriSNPr[30], and more recently Krisp[31]; however, PrimedSherlock requires primers designed using PrimedRPA[32] or alternative methods as input, CriSNPr was developed specifically for variant detection, and Krisp does not include PAM site specification, although these can be manually introduced into the primer sequence. Consequently, there is a need for a fully automated primer and gRNA design tool specific for nucleic acid detection that is scalable and requires minimal user intervention to help researchers overcome the initial hurdle in assay development.

Advancements in computational tools and the availability of pathogen genomic data provide an opportunity to identify alternative diagnostic markers from previously unexplored genomic regions. In this study, we present PathoGD, a bioinformatic pipeline developed to facilitate high-throughput and rapid design of highly specific RPA primers and gRNAs for CRISPR-Cas12a-based pathogen detection. To the best of our knowledge, this is the first bioinformatic tool providing a streamlined end-to-end workflow for primer and gRNA design in pathogens, thereby simplifying the process which normally requires multiple stages. PathoGD leverages available pathogen genomic data to accelerate the in silico aspect of CRISPR–Cas12a-based assay design and is scalable to large datasets. We demonstrate the application of PathoGD in designing highly specific primers and gRNAs for five clinically relevant bacterial pathogens and provide experimental validation of these designs for two common bacterial pathogens, *N. gonorrhoeae* and *Streptococcus pyogenes*. In addition, we compare the gRNA designs from PathoGD and PrimedSherlock for six viruses based on a common database of target and non-target genomes.

Results

PathoGD pipeline overview

PathoGD is an automated, command-line tool written in Bash and R for primer and CRISPR-Cas12a gRNA design that integrates the functionalities of several widely used bioinformatic tools (Supplementary Table 1). PathoGD incorporates two commonly used approaches in comparative genomics, pangenome, and *k*-mer analyses, either of which can be invoked based on user preferences. These approaches make up the two independent modules of the pipeline (Fig. 1, Supplementary Fig. 1), with multiple workflows available under each module. The pipeline takes as input the module and workflow selection, and a configuration file specifying multiple parameters, including: (i) target and non-target taxa; (ii) the National Center for Biotechnology (NCBI) database; (iii) genome assembly level; (iv) gRNA length, and (v) the desired threshold of gRNA prevalence across target genomes. Depending on the module and workflow selected, some of these parameters are mandatory while others are optional. Users can opt to download genomes automatically from the NCBI Assembly database (now NCBI Datasets)[33], or alternatively provide their own set of curated target and/or non-target genomes. Because certain species may be overrepresented in the database, a 'check' workflow is included which can be used to determine the number of genomes present prior to downloading genomes or running the pipeline. Genome subsampling can be selected if needed to reduce computational runtime for these overrepresented species.

PathoGD pangenome and k-mer modules

The pangenome and *k*-mer modules in PathoGD employ distinct methods to identify potential gRNA sequences within the target genome space. In the pangenome module, highly conserved protein-coding genes (≥ 90% prevalence across target genomes) across the entire genome are first identified as potential targets for gRNA and primer design. The protein sequences of these potential genes are clustered with protein sequences from non-target genomes to exclude genes that could potentially result in assay cross-reactivity (see Methods). The remaining potential genes are individually aligned, and canonical TTTN PAM sites are identified within each gene. The sequence downstream of the PAM site of a user-specified length *k* is stored as a potential gRNA sequence, or the gene is discarded if no PAM site is present. In genes where multiple PAM sites exist, ten gRNAs per PAM site are randomly selected, resulting in a maximum of 40 gRNAs per gene. A subsequent sequence comparison step eliminates gRNAs that are present in any non-target genome with up to two mismatches.

While the pangenome approach identifies universal genes, it excludes noncoding regions of the genome, including intergenic regions, ribosomal RNA, and other noncoding RNA. These non-translated regions make up an average of 12% of the bacterial genome[34,35], comprise many regulatory elements with key functions[36] and potentially contain species-specific sequence signatures[37] that could be exploited for diagnostics. To this end, we developed the *k*-mer module, a gene-agnostic approach interrogating both coding and non-coding regions across the entire genome. In the *k*-mer module, all sequences of a user-specified length *k* in both target and non-target species, including the human genome, are enumerated. Sequences common between the target and non-target genomes are removed, allowing for up to two mismatches in the non-target genomes (see Methods). A second *k*-mer enumeration step is performed to retain only sequences that are unique to the target species, downstream of a PAM site, and present above a user-specified genome prevalence threshold as potential gRNAs.

In both modules, gRNAs with the potential to form hairpin structures are eliminated, and upstream and downstream sequences flanking the remaining gRNAs are extracted for primer design. The primer design step is optional in the *k*-mer module and can be excluded where pre-amplification is not required for the assay. An in silico PCR is subsequently performed against all target and non-target genomes to estimate the sensitivity and specificity of each primer set. The output of PathoGD is a tab-delimited file providing up to five matching RPA primer pairs for each gRNA, along with associated information including GC content, expected amplicon size, and potential cross-reactivity with off-target organisms. This comprehensive output provides users with the flexibility to apply further filtering based on criteria such as primer and gRNA prevalence across target genomes, average copy number, and potential off-target hits.

PathoGD was designed primarily for use with RPA and Cas12a-based applications. The primers and gRNAs designed by default can theoretically be adapted for Cas13a-based applications by adding the T7 promoter sequence to the appropriate primer, and this information is included in the output file. To increase flexibility, we have incorporated an option for designing guide RNAs without PAM requirements, which can help generate more potential targets. Additionally, there is an option for designing primers specifically for PCR applications to accommodate users who prefer conventional amplification methods over isothermal approaches.

Minimizing potential off-target activity

Off-target activity in CRISPR-based diagnostics can occur through a combination of non-specific activity at the RPA and CRISPR–Cas stages of the assay. Non-specific amplification can occur as RPA has been shown to tolerate up to nine primer-template mismatches[38–40]. Combined with the potential imperfect binding of gRNA to non-target amplicons[41] or background DNA, this may result in the activation of trans-cleavage and generation of off-target signal, thereby reducing assay specificity.

PathoGD aims to minimize the possibility of false positives at the sequence level by implementing multiple steps to identify potential primer or gRNA binding sites in non-target genomes. In the pangenome module, core genes that are highly similar between target and non-target species are excluded through the clustering of target and non-target protein sequences at a 90% average amino acid identity (AAI) threshold. Cas12a has been demonstrated to tolerate mutations in the PAM and partial mismatches between gRNA and target region which permit target binding and cleavage to varying degrees[30,41]. To account for possible trans-cleavage activity resulting from imperfect target binding, candidate gRNAs with less than three mismatches to any sequence in the non-target or human genome databases are eliminated through a *k*-mer comparison step (see Methods). These steps refine the pool of gRNAs to those with minimal off-target effects before proceeding to primer design, reserving computational resources for only promising and specific candidates. Additionally, mismatches between the gRNA and target DNA that occur in the seed region- the first 6–7 bases downstream of the PAM site, are more deleterious to Cas12a[42,43] compared to mismatches at PAM-distal positions. To distinguish between these two categories, PathoGD incorporates a separate field denoting gRNAs where mismatches to non-target DNA occur exclusively outside of the seed region to allow for the selection of more robust gRNAs where available.

The selection of a primer and gRNA combination often involves a compromise between sensitivity and specificity. To facilitate decision-making, additional information is provided in the output: (i) the proportion of genomes estimated to amplify from each primer set across the target and non-target genome databases, allowing up to two and eight overall mismatches for the target and non-target genomes, respectively, (ii) prevalence of gRNAs with no mismatches in the target genome database, (iii) gRNAs whose mismatches occurred exclusively at PAM-distal positions where they are mostly tolerated (see Methods). We suggest the following criteria be used where possible for identifying a highly reliable set of primer and gRNA: ≥90% prevalence of primers and gRNAs in the target genome database, exclusion of primer sets with up to eight mismatches to off-target genomes, exclusion of gRNAs where mismatches to non-target genomes occur exclusively at PAM-distal positions, and prioritization of TTTA, TTTC, and TTTG PAM sites, due to demonstrated lower activity of wild-type LbCas12a nuclease on TTTT PAM sites[44,45], which we also confirmed (Supplementary Fig. 2). We note that the above is intended as general guidelines and should be adapted for different organisms or specific experimental contexts where necessary.

Computational performance of PathoGD for primer and gRNA design for five clinically relevant human pathogens

We used PathoGD to design primer and gRNAs for five pathogens of global significance, which collectively represent four different phyla and include both Gram-negative and Gram-positive bacteria: *N. gonorrhoeae, Treponema pallidum, Chlamydia trachomatis, S. aureus*, and *S. pyogenes*. These pathogens are high-priority global threats, according to the World Health Organization (WHO) with urgent benchmarks in place for novel diagnostics[46], urgent vaccine development[47], and new antibiotics[48] for pathogen elimination. Primers and gRNAs were designed for each species using genomes obtained from the NCBI Assembly database (Supplementary Table 2), subsampled to 100 genomes each except for *T. pallidum* where 92 genomes were available. The designs were subsequently validated against a larger genome database for each species (see Methods), containing sequences with global representation and spanning a period of nearly two decades.

To improve specificity for primer and gRNA design, we included both 'target' and 'non-target' genomes from species belonging to the same genus as the target species (Supplementary Data 1), as well as the human genome (see Methods). All designed primers and gRNAs were subsequently validated against the complete, non-subsampled 'target' and 'non-target' database for each species. Each run was performed on a computational cluster comprising Intel(R) Xeon(R) Gold 6342 CPU @ 2.80 GHz (Icelake) processors using 32 CPU and 64GB RAM. Excluding the genome download step, run times ranged from 44 min to 5 h 26 min for the pangenome module, and 41 min to 3 h 15 min for the *k*-mer module, and were largely correlated with the size of the target and non-target genome databases (Supplementary Data 2). Pipeline bottlenecks occurred at four stages:

alignment of target genes or amplicons; enumeration of *k*-mers for non-target genomes; evaluation of guide specificity, and target gene annotation (specific to the pangenome module). Across the five species, these three steps accounted for 74.2 ± 5% and 79.1 ± 15% (mean ± standard deviation (SD)) of the total run times for the pangenome and *k*-mer modules, respectively. The computational runtime of PathoGD was significantly improved by the incorporation of tools employing efficient algorithms, including GenomeTester4[49] for enumerating *k*-mers and manipulating the resulting lists, Primer3[50] for primer design and evaluation, and isPcr[51] for in silico PCR amplification.

Identification of primers and gRNAs using PathoGD

The number of target-specific gRNAs identified by PathoGD varied across the five species (Fig. 1a, c). Despite having the largest genome sizes among the five species, *S. aureus* (~2.83 Mbp) and *N. gonorrhoeae* (~2.15 Mbp) had the least number of identified gRNAs across both pangenome and *k*-mer modules—1017 for *S. aureus* and 409 for *N. gonorrhoeae* (Fig. 2a, c; Supplementary Fig. 3; Supplementary Data 2 and 4). In contrast, the two species with the smallest average genome sizes (~1.1 Mbp) had the highest numbers of gRNAs—10,035 for *C. trachomatis* and 15,157 for *T. pallidum*, respectively (Fig. 2a, c; Supplementary Fig. 3; Supplementary Data 2 and 4), consistent with their relatively lower genetic diversity and highly adapted lifestyles[52–54]. Overall, 5317 gRNAs were identified for *S. pyogenes* from a smaller number of core genes compared to *S. aureus* and *N. gonorrhoeae*. These results translate to a higher number of actual primer/gRNA combinations available, given that up to 5 primer pairs were reported for each gRNA. Additionally, the above results were derived from observations where gRNAs and their associated primer pairs exhibited perfect matches to ≥90% of genomes in the complete target database, underscoring the utility of PathoGD in high-throughput identification of potential assay targets.

Only partial overlaps of gRNAs were observed between the pangenome and *k*-mer modules (Fig. 2c) due to the different algorithms employed by each module at the initial candidate gRNA identification steps (see Methods). Specifically, in the pangenome module, heuristics employed for narrowing down the initial search space, including core gene elimination via average AAI to non-target genes and subsampling of a maximum of 10 gRNAs per canonical PAM for each gene, mean that certain gRNA candidates will be missed. In the *k*-mer module, *k*-mers are shortlisted only if present above a certain genome prevalence threshold, and if their flanking regions are sufficiently conserved. For the *k*-mer module, the percentage of gRNAs belonging to or overlapping with noncoding regions was estimated to range from 9.7 to 53.6% for the five species tested (Fig. 2d). These results, along with the substantial proportions of gRNAs unique to each module, support the complementarity of the pangenome and *k*-mer approaches in designing CRISPR–Cas12a diagnostic assays.

A further distinction between pangenome and *k*-mer modules in PathoGD is that the latter allows for the identification of target-specific sequences that occur in multiple copies throughout the genome. These sequences are attractive candidates for diagnostic targets in samples where low pathogen abundance is an issue, due to their potential for enhancing assay sensitivity. Additionally, the risk of false negatives is minimized in cases where certain copies are absent or have undergone mutations. The use of multi-copy markers as targets in molecular assays is not uncommon, for example IS2404 in *Mycobacterium ulcerans*[55], IS481 in *Bordetella pertussis*[56] and IS711 in *Brucella* spp.[57]. As part of the *k*-mer module, multi-copy target-specific sequences are identified in an intermediate step of the pipeline and are the final output for a workflow not requiring primer design (Supplementary Fig. 1). We identified 32 gRNAs with an average of ≥2 copies across *T. pallidum* (24 gRNAs), *N. gonorrhoeae* (4 gRNAs) and *S. pyogenes* (4 gRNAs) in our dataset (Fig. 2b, Supplementary Data 4); however, conserved primer binding sites were identified for only 14 of these (all in *T. pallidum*) likely due to polymorphisms in the gRNA flanking regions. The actual copy numbers of these individual gRNAs are highly variable within strains of a species (Supplementary Fig. 4) and results will likely differ when using a different subset of genomes.

Comparison of PathoGD to PrimedSherlock

Currently, there is no reported pipeline for designing both primers and gRNAs for CRISPR–Cas12a diagnostic applications. However, a previously published tool, PrimedSherlock[29], described the design of gRNAs using primers generated from PrimedRPA[32], although published data are, to date, only available for viral genomes. Using the input target and off-target genome sequences reported in PrimedSherlock, we designed primers and 20-nucleotide gRNAs for 3 species of *Flavivirus*—West Nile virus (WNV), Zika virus (ZIKV), and four serotypes of Dengue virus (DENV-1, DENV-2, DENV-3, and DENV-4) using the PathoGD *k*-mer module, without genome subsampling. We compared the results of PathoGD to that of PrimedRPA and PrimedSherlock in terms of the potential to amplify their genomic regions and the prevalence and conservation of primer and gRNA binding sites across target and off-target genomes calculated using the same approach (see Methods).

The small average genome sizes (~11 kb) of these flaviviruses translated to few gRNAs, with less than 15 gRNAs obtained for DENV-1, DENV-2, DENV-4 and ZIKV, and 21 and 83 gRNAs for DENV-3 and WNV, respectively (Supplementary Fig. 5a). Across all viruses, gRNAs from PathoGD had comparable or higher prevalence and conservation across the target genomes compared to PrimedSherlock, both when considering perfect matches and when allowing at most one mismatch between the gRNA and target genomes (Supplementary Fig. 5b). Unsurprisingly, the prevalence of designed gRNAs was lower in these viruses compared to bacterial genomes, due to their high genetic diversity (Supplementary Fig. 5b). Only two gRNAs were common between PathoGD and PrimedSherlock- one for DENV-3 and one for WNV, present across ~92% genomes in their respective databases. All gRNAs were specific to their target viruses, with only one DENV-1 gRNA from PrimedSherlock matching DENV-3 genomes at all but one position at the 3' end. The primer binding sites were less conserved for dengue viruses compared to ZIKV and WNV, with no PathoGD primer candidates present in more than 50% of target genomes for DENV-1 and DENV-2 at a strict threshold of no allowed mismatches (Supplementary Fig. 6a). However, allowing for mismatches across the forward and reverse primers increased the number of available primer candidates (Supplementary Fig. 6a) without compromising in silico primer specificity (Supplementary Fig. 6b). Although the DENV-3 and DENV-4 primers reported in PrimedSherlock were more highly conserved (Supplementary Figs. 6a and 7a), PathoGD candidates had better overall prevalence and conservation for all six viruses when considering the primers and gRNAs in combination (Supplementary Fig. 7).

Experimental validation of primers and guide RNAs

To experimentally validate the primers and gRNAs designed in silico, we performed RPA–CRISPR–Cas12a assays for *S. pyogenes* and *N. gonorrhoeae*, following a previously established protocol[13]. Seven sets of primer/gRNA candidates from *S. pyogenes* and six from *N. gonorrhoeae* present in ≥90% of genomes from their respective databases were selected from the list of primer and gRNAs generated from the PathoGD pangenome module (Fig. 3a, Fig. 4a, Table 1, Supplementary Data 3 and 4). Additionally, primers and guides were manually designed for the widely used *N. gonorrhoeae* diagnostic target *porA* (Table 1, Supplementary Data 3).

All seven primer/gRNA sets detected *S. pyogenes* at 10,000 copies/μL and exhibited no cross-reactivity with genomic DNA of other species belonging to the same genus. These include the closely related *S. dysgalactiae* and *S. agalactiae*, as well as *S. anginosus* and *S. pneumoniae* (Supplementary Table 2, Fig. 3b, Supplementary Data 4). In all instances, the fluorescence signals plateaued within the first 5 min, suggesting maximal reporter cleavage efficiency by Cas12a for the starting DNA concentrations used. A similar result was obtained for *N. gonorrhoeae*, where all six primer/gRNA sets accurately detected *N. gonorrhoeae* at 10,000 copies/μL. Only a minimal background fluorescent signal was observed for a panel of closely related species, including *N. meningitidis* and *N. lactamica* (Fig. 4b, Supplementary Data 4). In one of three runs, a

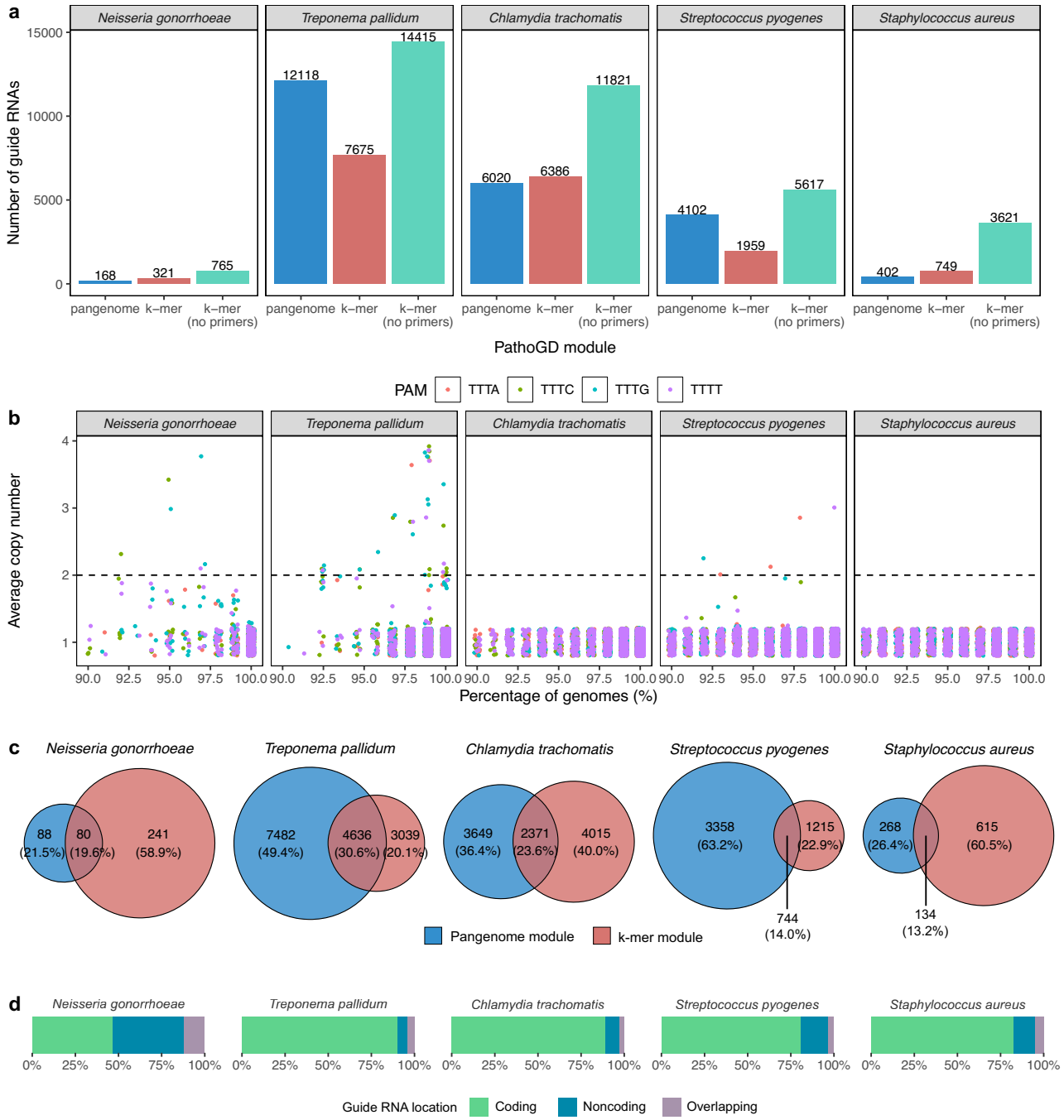


Fig. 2 | PathoGD in silico results of pangenome and k-mer modules for five bacterial species. a Number of target-specific guide RNAs identified from the pangenome and *k*-mer modules in PathoGD. The *k*-mer (no primers) module refers to the workflow without primer design (see Supplementary Fig. 1). Results for the pangenome and *k*-mer modules were filtered to only observations where gRNAs and primers were present across 90% of strains in each target species, primers had at least nine mismatches to the non-target species, and at least one mismatch of gRNA to non-target genomes occur in the PAM or seed region. **b** Average copy numbers of gRNAs identified from the *k*-mer (no primer) module for the five bacterial species. Black dotted lines refer to the cut-off at 2 copies. **c** Unique and overlapping gRNAs identified from the pangenome and *k*-mer modules in (**a**) for five bacterial species. **d** Distribution of gRNAs designed with the *k*-mer module across the coding, non-coding, or overlapping regions within the genome. All results were generated using the 'user_all_subsample' workflow, with 100 genomes randomly subsampled for primer and guide RNA design for each target species, except for *T. pallidum* ($n = 92$). The subsampled genomes used in both modules are identical for each target species.

minor fluorescence signal was observed in *Candida albicans* for the *porA* gRNA 20 min into the CRISPR/Cas12a assay, suggesting a possible contamination event (Fig. 4b). For both organisms, an assessment of analytical sensitivity revealed clear differences in the performance of primer/gRNA sets, especially at lower pathogen concentrations (Figs. 3c and 4c). Together, these results highlight the robustness in the specificity of the primers and gRNAs designed by PathoGD, as well as the need for evaluating multiple candidates to identify the targets with high sensitivities.

In addition, we evaluated the potential of using multi-copy gRNAs for amplification-free detection of *N. gonorrhoeae*. Our amplification-free assay was based on an input gDNA of 250 ng/μL, determined from previous experimental evaluation to be the minimum amount needed for the detectable signal above background levels. From an initial assessment of 6

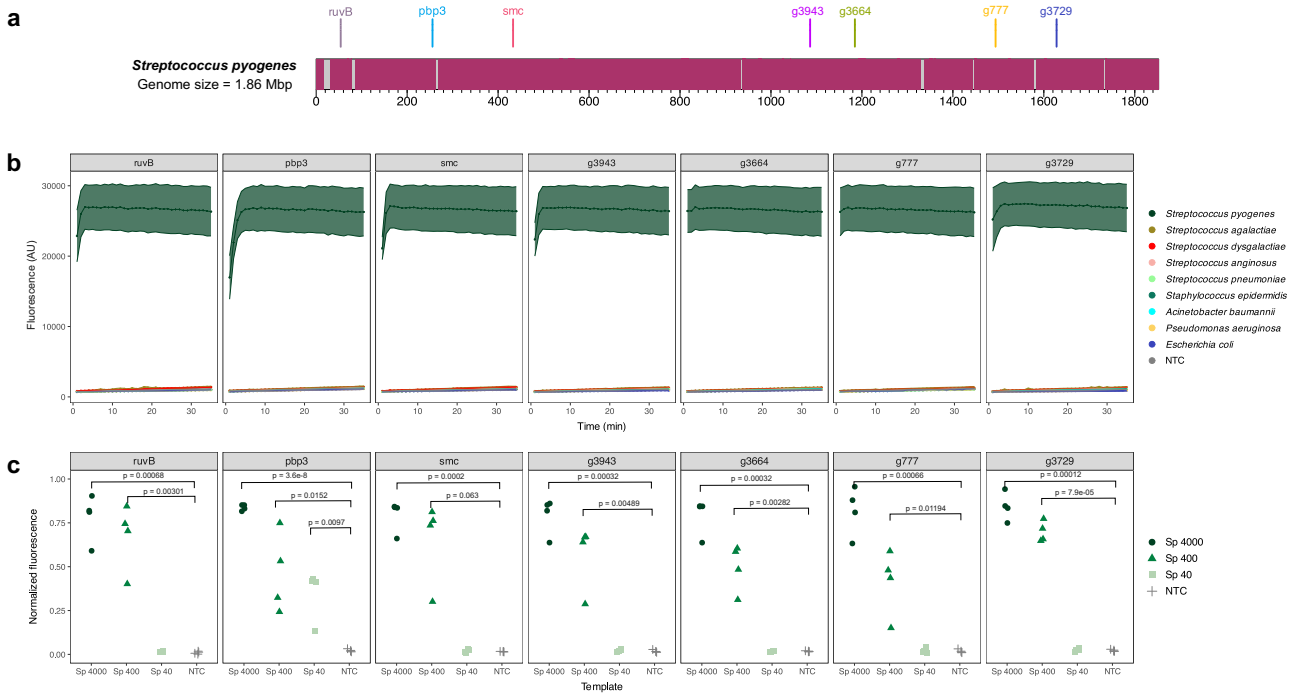


Fig. 3 | Experimental validation of primers and gRNAs for *Streptococcus pyogenes* in an RPA/CRISPR–Cas12a assay. a Distribution of guide RNAs across the *S. pyogenes* ATCC 19615 genome (RefSeq assembly accession GCF_000743015.1). **b** Fluorescence profiles of guide RNAs across 35 min for the *S. pyogenes* ATCC 19615 strain. Data represent the average fluorescence values of three experimental replicates. Error bars represent the standard error of the mean (SEM). The specificity of each gRNA was tested against a panel of closely related organisms and other organisms likely to be present at the site of infection. **c** Limit of detection of the guide RNAs using 4000, 400, and 40 copies/μL of *S. pyogenes* gDNA, with NTC as the negative control. Data represent the normalized fluorescence at the assay endpoint for four replicates. Samples were compared to the NTC using a one-sided *t*-test.

multi-copy gRNAs (average copy number ≥1.5) present in ≥90% of the target genomes (Fig. 5a, Supplementary Data 3), four non-overlapping candidates with the highest fluorescent signals (Fig. 5b) were further screened in an amplification-free CRISPR–Cas12a assay. These gRNAs were assessed individually and incrementally pooled to generate 11 possible pooled combinations (Fig. 5c, Supplementary Data 4). These four gRNAs are present in 3–9 copies within the FA1090 genome and cumulatively target 22 regions when pooled in a single reaction (Fig. 5a, b).

All multi-copy gRNAs successfully detected *N. gonorrhoeae*, albeit with a much slower signal accumulation compared to an assay with RPA preamplification. As expected, combinations of gRNAs with higher cumulative copy numbers resulted in higher endpoint fluorescence, with the signal in most combinations being driven by g686, the best-performing gRNA (Fig. 5b, c). Pooled multi-copy gRNAs resulted in at least a 1.5-fold increase in fluorescence signals at the assay endpoint of 58 min compared to *porA*, a single copy marker (Supplementary Figs. 8 and 9). The fluorescence signals did not plateau for any of the gRNA sets (Supplementary Fig. 8), even when the assay was extended to 118 min (Supplementary Fig. 8), consistent with previous observations[58] and suggesting suboptimal activation of Cas12a in a preamplification-free assay. Encouragingly, however, multi-copy gRNAs were highly specific, with only minimal off-target signal observed in the presence of four pooled gRNAs beginning at 20 min (Supplementary Figs. 8 and 9).

Discussion

Critical to the success of CRISPR-based diagnostics is the design of highly sensitive and specific primers and gRNAs for target detection, a process generally involving multiple iterative steps. Here, we describe PathoGD, an end-to-end bioinformatic pipeline for the high-throughput and rapid design of highly conserved and specific primers and gRNAs for CRISPR–Cas12a diagnostic applications. PathoGD integrates multiple tools employing efficient algorithms and requires only a user-specified configuration file as input, significantly reducing the complexity and time required for assay design. PathoGD is reliant on a well-curated target (inclusivity) and non-target (exclusivity) genome database to capture the full spectrum of genetic diversity within the target species while excluding closely related species and organisms likely to co-occur at the site of infection. We demonstrated the utility of PathoGD for designing primers and gRNAs for five diverse species of bacterial pathogens. Computational runtimes were largely dependent on database sizes, with the longest runtime observed for *S. pyogenes* (5 h 26 min) for the pangenome module; however, all other runs were completed within 4 h for both modules. Overall, the use of a small subset of 100 genomes per species was sufficient to generate an extensive set of potential targets that were valid across the larger complete target database.

PathoGD eliminates the reliance on traditionally used marker genes when selecting a target region for CRISPR-based detection by leveraging publicly available genome assemblies. In addition to a gene-centric (pan-genome) approach, the incorporation of a complementary gene-agnostic (*k*-mer) approach allows for a systematic and unbiased exploration of sequences in both protein-coding and non-translated regions of the genome for more challenging organisms. As the field of molecular diagnostics is transitioning towards a dual-target approach to minimize the risk of false negatives, an essential step is the evaluation of multiple candidate primers and gRNAs as their sensitivities and specificities in vitro may vary. The comprehensive output generated by PathoGD provides flexibility in the selection of candidates for experimental validation. Equally important is the continuous monitoring of newly sequenced genomes of pathogens for potential mutations that could result in diagnostic escape. The scalability of PathoGD ensures that the design process can easily be revisited with the newly sequenced genomes included, or alternatively, previously designed primers and gRNAs can be evaluated against new genomes to validate their ongoing relevance.

The number of target-specific gRNAs identified by PathoGD varied across species and could be attributed to the following factors: core genome

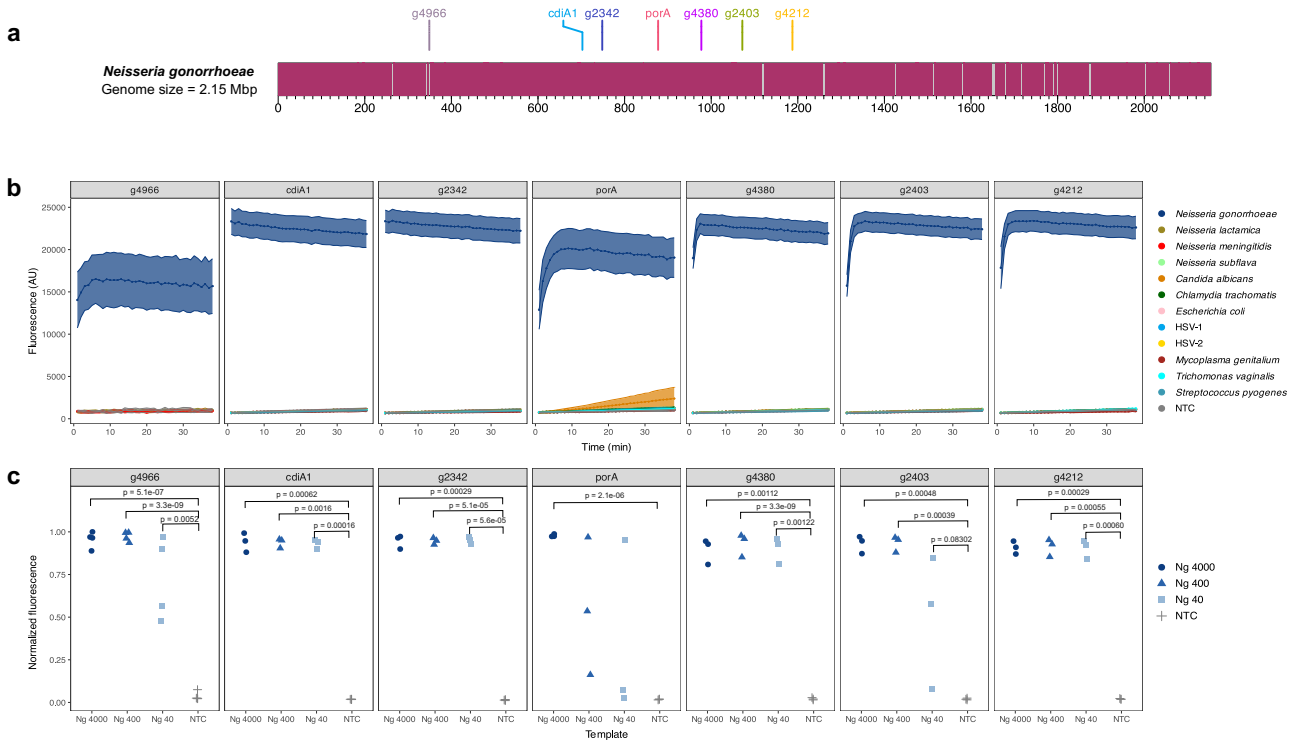


Fig. 4 | Experimental validation of primers and gRNAs for *Neisseria gonorrhoeae* in an RPA/CRISPR–Cas12a assay. a Distribution of guide RNAs across the *N. gonorrhoeae* FA1090 genome (RefSeq assembly accession GCF_000006845.1). **b** Fluorescence profiles of guide RNAs across 38 min for *N. gonorrhoeae* FA1090 strain. Data represent the average fluorescence values of three experimental replicates. Error bars represent the SEM. The specificity of each gRNA was tested against a panel of closely related organisms and other organisms likely to be present at the site of infection. **c** Limit of detection of the guide RNAs using 4000, 400, and 40 copies/µL of *N. gonorrhoeae* gDNA, with NTC as the negative control. Data represent the normalized fluorescence at the assay endpoint for at least three experimental repeats ($n = 4$ for g4966, $n = 3$ for all other gRNA sets). Samples were compared to the NTC using a one-sided *t*-test.

size, allelic variants within core genes, evolutionary relationship with closely related species, and the genetic nature and diversity of strains subsampled for each species. Organisms sharing large homologous regions with closely related species and frequently engaging in horizontal gene transfer present a challenge for target-specific primer and gRNA design. In our dataset of five species, this was exemplified by *N. gonorrhoeae* which shares between 80 and 90% of its genome sequence with its closest relative *N. meningitidis*[59], is naturally competent for DNA uptake[60,61], undergoes frequent recombination with other *Neisseria* species[62], and exhibits significant genetic variation between subtypes[63]. These factors contribute to complexity in marker selection for accurate and specific detection of *N. gonorrhoeae*, especially for specimens from extragenital sites where commensal *Neisseria* species are commonly found[63]. Indeed, commercial gonococcal nucleic acid amplification tests (NAAT) have displayed variable cross-reactivity with other *Neisseria* species and often require a supplementary assay for confirmation[64]. Across both pangenome and *k*-mer modules, PathoGD identified a total of 409 *N. gonorrhoeae*-specific gRNAs from 31 genes. Our experimental validation on a subset of these genes supports their potential as alternative targets for specific detection of *N. gonorrhoeae* that could be assessed for increased sensitivity at low pathogen concentrations, compared to *porA* which has been utilized by *N. gonorrhoeae* CRISPR–Cas assays to date[18,65,66].

Despite its smaller proportion of core genes compared to *N. gonorrhoeae*, we identified a substantially larger number of gRNAs ($n = 5317$) in *S. pyogenes* due to a higher number of species-specific genes. Unlike *porA* in *N. gonorrhoeae*, there is limited data regarding a widely used NAAT diagnostic target for *S. pyogenes*. We validated a subset of seven primers and gRNA combinations for the detection of *S. pyogenes* and demonstrated high specificity, highlighting the reliability of primers and gRNAs designed by PathoGD. Although only a single strain of target and non-target species was used for all validation assays, in silico evaluation against a large database of closely related species suggested that the gRNAs are robust for discriminating target and non-target organisms with a broad range of genetic diversity.

Across other assessed species, varying numbers of gRNAs were identified as a product of their genetic diversity, evolutionary history, ecological niche, and degree of relatedness to other species. Importantly, the targets we identified were validated across comprehensive databases with sequences that have wide geographical distribution and have persisted over evolutionary time scales, translating into increased longevity and robustness of the diagnostic assay. In our example runs the off-target databases consisted only of species belonging to the same genus as the target pathogen, which may be overly conservative considering that many may not be present in the same sites or environment as the target pathogen. Likewise, the inclusion of more distantly related species in the off-target database should be carefully considered for organisms where horizontal gene exchange across genera or higher taxonomic ranks is prevalent.

The lack of overlap between gRNAs generated from PathoGD and PrimedSherlock can be attributed to their different underlying approaches. Specifically, PathoGD starts by identifying potential gRNAs that serve as anchors for primer design, while PrimedSherlock designs gRNAs from predefined genomic regions amplified by user-provided primer sequences. Additional considerations in PathoGD that were absent in PrimedSherlock include the elimination of gRNA candidates with homopolymers of five or more nucleotides, as well as those matching the human genome. These highlight the currently evolving field of CRISPR diagnostics where no consensus exists yet for standard approaches in primer and gRNA design. Our results with the DENV-3 dataset suggested that some pipeline parameters may have to be relaxed when designing assays for viruses to increase the number of available primer and gRNA candidates, especially for rapidly mutating RNA viruses. The simultaneous design of primers and gRNAs is also advantageous for the upfront elimination of genomic regions that do not contain PAM sites or are poorly conserved across the target species.

Table 1 | Primer and guide RNA sequences used for experimental validation on *Neisseria gonorrhoeae* and *Streptococcus pyogenes* gDNA

Species	gRNA name	PathoGD module	Forward primer	Reverse primer	gRNA spacer
Neisseria gonorrhoeae	cdiA1	Pangenome	5'-AAGAAAGTCTTCAGCGAAAATGGTAAGTTG	5'-CAAAGTATAATTTTGTTCACGATGCCCTGTTT	5'-ATCATGTCCTTTACGACGCGCA
	g4966	Pangenome	5'-ATAGCTATGTTTCAACTTGTGAATTTGTAG	5'-CAGATATAATCAAAATCTTCTTGTTGGTCA	5'-GCCTTTGGGTTTACGTCCATAA
	g2342	Pangenome	5'-GAATTGCAATTGGTTTCACACAAAGATGAG	5'-TATCCAGCATCGCATAGCCTGATTCCAAGTTT	5'-ACGTGGAACGAGTGCGTAGGCC
	g2403	Pangenome	5'-CACTATCACATCGGCATTCCCAAATACAACG	5'-TTATTTGGGAATATCGGGCAGGATAAAAGG	5'-ATAAAGCCGTCGCAAAGGGTAT
	g4212	Pangenome	5'-GTTAAATTTTTCAGGCGGTTCATTGGATTG	5'-TATTTCTTTTCCCGAATCAATTTGTTGTAACG	5'-CGGCTTTTCTTGAAGTTGATAA
	g4380	Pangenome	5'-GAACAACTTTTAAACCACGCCGTCCGCTATG	5'-TTTCGTGTGCCAGTTCTGCCAGCCCTCAATTT	5'-GATTTCGAGCTTTCGGTAGTAT
	porA	Pangenome	5'-AATAATAATGTGGCTTCGCAATTGGGTATT	5'-GCATATTTAAGGGCATAATTTCCGAAAAAG	5'-ACTCGGAACAAATTGAATGCTG
Streptococcus pyogenes	g777	Pangenome	5'-CAAGTAAGTTTAATGATGATAGTGCTATTGTC	5'-TACCACTACCAGTAGTATCAGTAAAATAAACA	5'-ATTGACCTCGTGAAAACTTAGC
	pbp3	Pangenome	5'-CCATTTACAGTTATAATTACATGCTTAAAGGC	5'-GTTGTTTTAAATATAGCTAAATCATCCTCGTG	5'-ATTTTCGACACTAGTAGCTACA
	g3729	Pangenome	5'-TACTATATACTAAAGGCTGTATGCGAATCAAT	5'-TATCTGTTAAGGTAACCACATAGGCTCTTTTA	5'-GTACAACAGTCATATTCCCCGA
	g3664	Pangenome	5'-CTTAAATAGTGAACAAAAAGCTGTATCTGCTA	5'-TAACCTTTTGGAGTCGAAATAATATCAATCTG	5'-GGTAAGTTATTAGGAGTGAGTC
	g3943	Pangenome	5'-GAAGAGATAGTAGGTTTTTCAACTCAAGAATA	5'-GTAAGCTTCAACTAATTTAGGATAGACATCTT	5'-AGACGGTTAAGCATCTTATACA
	smc	Pangenome	5'-GAAAAGAAATTTGAGCAAGCTAATCAATCTC	5'-CATAGTCCTCTATTTCAAATCTTAATTGGACT	5'-CCACTGGGGTAGTCTTTGAATG
	ruvB	Pangenome	5'-CGAATATATTGGACAAGATAAAGTTAAGGAAC	5'-GGTCCAAATAATAAAACATGGTCTAAAGATTC	5'-TAGAAGCAGCTAAAAGACGTGA
Neisseria gonorrhoeae	g686	*k*-mer	N/A	N/A	5'-AAATTTAAGGCCTAATTTGCCT
	g721	*k*-mer	N/A	N/A	5'-CCGAACAAGCGGTCCGGACTCC
	g716	*k*-mer	N/A	N/A	5'-CCACTTATCCGCGCTTTCGCCA
	g752	*k*-mer	N/A	N/A	5'-CACCAACGCCTGCTGTATAGCG
	g588	*k*-mer	N/A	N/A	5'-ACCGGGCGGACAGCCGGTTCCG
	g687	*k*-mer	N/A	N/A	5'-AACTTTTCCAGTGCGGCGAGAT

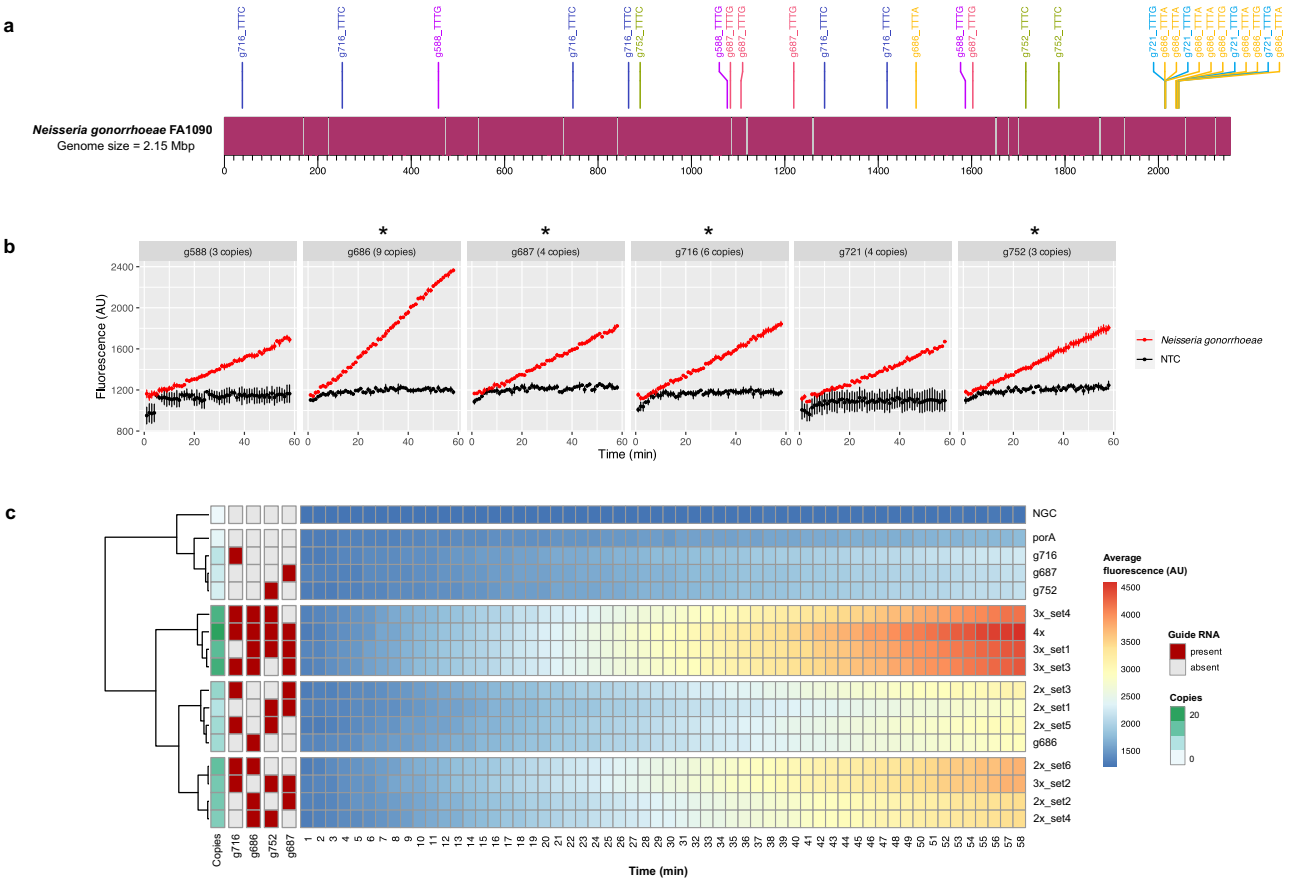


Fig. 5 | Experimental validation of gRNAs designed using *k*-mer module for *Neisseria gonorrhoeae*. **a** Distribution of six selected multi-copy gRNAs across the *N. gonorrhoeae* FA1090 genome. Guide RNAs are present in multiple copies throughout the genome and are denoted by their names and PAM sequence separated by an underscore. **b** Initial assessment of six multi-copy gRNAs on *N. gonorrhoeae* FA1090 gDNA in an amplification-free CRISPR–Cas12a assay. The copy number for each gRNA in the FA1090 genome is listed next to the gRNA names. Data represent average fluorescence signal across 58 min. Error bars represent SEM across three experimental repeats (two technical replicates for *N. gonorrhoeae* in each experimental repeat). The four gRNAs selected for pooling are indicated with an asterisk. **c** Heatmap of average fluorescence profiles for individual and pooled multi-copy gRNAs in an amplification-free CRISPR–Cas12a assay using four selected gRNAs. NGC refers to the control with no guide RNA added. Fluorescence values were averaged across multiple experimental repeats ($n = 3$ for porA, $n = 6$ for all other gRNA sets). The presence–absence profiles of individual gRNAs across all individual and pooled combinations tested and their cumulative copy numbers in the FA1090 genome are annotated on the left of the heatmap.

Previous studies have shown success in preamplification-free detection of SARS-CoV-2[58] and hepatitis C virus and *Mycobacterium ulcerans* with Cas13a and Cas12a, respectively[67]. The availability of candidate gRNAs present in multiple copies across the genome in *N. gonorrhoeae* presented an opportunity to evaluate the potential increase in detection sensitivity through the use of multi-copy gRNAs, both individually and pooled. Our experiment supported previous studies reporting an increased sensitivity with the inclusion of more gRNAs, with a clear relationship observed between signal accumulation and cumulative regions targeted by the gRNAs. While not sufficient to detect pathogens at low concentrations on its own, the pooling of highly specific multi-copy non-overlapping gRNAs may be valuable when used in conjunction with post signal amplification approaches compatible with CRISPR–Cas assays[68,69]. The presence of multi-copy species-specific gRNAs in bacterial genomes is not ubiquitous, as demonstrated by the in silico results for the five species, but can be exploited where available as the CRISPR diagnostics field shifts towards amplification-free methods for nucleic acid detection. In this regard, PathoGD can address the need for rapid identification of highly specific multi-copy potential diagnostic targets.

There are several limitations of PathoGD, some of which can be addressed in future versions of the tool. First, gRNA designs in PathoGD are limited to those adjacent to Cas12a-preferred TTTN PAM sites or do not require an adjacent PAM site. Future improvements could expand its utility for other Cas nucleases with different PAM requirements. Second, the success of a run is dependent on the taxonomic accuracy of genomes in the target and non-target databases. Single contamination in either database due to taxonomic or assembly anomalies is sufficient to remove or significantly reduce the number of potential target-specific gRNAs identified. This is due to the *k*-mer comparison step employed by PathoGD, where sequences present in both databases would be eliminated. A future version would benefit from an integrated taxonomic validation step. Finally, PathoGD refrains from ranking or scoring the designed primers and gRNAs, as their binding and cleavage efficiencies are dependent on multiple factors that may not be fully captured by a simple scoring system. Studies focused on CRISPR editing applications in eukaryotes have shown that gRNA sequence composition, chromatin accessibility, RNA secondary structure, and free energy contribute to gRNA efficiency[70]. Future large-scale studies investigating the above factors could pave the way for incorporating machine learning in predicting gRNAs with maximal activity for bacterial diagnostics, such as that implemented with Cas13 for viruses[71].

Methods

Pangenome module

To identify a set of targets for amplification, gene annotation was performed using Prokka (v1.14.6)[72] for all target or subsampled target genomes, depending on user input. The pangenome was estimated using Roary

(v3.13.0)[73] for the same set of target genomes. Conserved genes present in at least 90% of all genomes were identified as potential targets for amplification. Structural gene annotation was performed for all non-target genomes using Prodigal (v2.6.3)[74]. Protein sequences of conserved target and non-target genomes were clustered at 90% amino acid identity threshold using cd-hit (v4.8.1)[75], and only gene families from target genomes that did not cluster with genes from non-target genomes were retained as target-specific genes. All target-specific gene families were individually aligned using MAFFT v7.505[76], and a consensus gene alignment was obtained using EMBOSS (v6.6.0; http://emboss.open-bio.org) as the template for gRNA and primer design. Canonical PAM sites (TTTA, TTTC, TTTG, TTTT) were identified from each consensus gene alignment on both sense and antisense strands. Sequences downstream of the PAM site of a user-specified length *k* were stored as potential gRNA sequences, or the gene was discarded if no PAM sites were present. Up to 10 potential gRNA sequences were retained per PAM site to give a maximum (31) of 40 gRNA sequences per gene. Low-quality potential gRNAs were removed based on their propensity to form hairpin structures using the 'check_primers' task from Primer3 (v2.6.1)[50]. A *k*-mer comparison was performed between the remaining potential gRNA sequences and *k*-mers of the same length enumerated from non-target genomes using GenomeTester4 (v4.0)[49] to eliminate gRNAs present in the non-target genomes, allowing at least 2 mismatches. Flanking sequences of up to 150 bp (where available) were extracted for the remaining gRNA sequences using SeqKit (v2.3.1)[77], and primers designed for an amplicon size range of 90–300 bp using Primer3. A maximum of five primer pairs were designed for each gRNA. An in silico PCR was performed against all target and non-target genomes with the designed primers using isPcr (v33)[51] to estimate the prevalence of each target amplicon across the target and non-target genome databases. Similarly, all gRNAs were mapped to target and non-target genomes using BBMap (v38.96)[78] to obtain an estimate of gRNA prevalence.

k-mer module

k-mers of a user-specified length were enumerated for all target or subsampled target genomes and all non-target genomes using GenomeTester4, followed by a *k*-mer comparison between the target and non-target genomes to eliminate potential gRNA sequences with up to two mismatches in the non-target genomes. A second *k*-mer enumeration step of length $k + 4$ was performed for the target genomes to identify target-specific *k*-mers with PAM sites. Low-quality potential gRNAs from this list were removed based on their propensity to form hairpin structures using the 'check_primers' task in Primer3. The target-specific *k*-mers were mapped to the target genomes using BBMap and flanking sequences of up to 150 bp (where available) were extracted for each *k*-mer as potential amplicons. All amplicons were clustered using cd-hit, and clusters above 90% nucleotide identity were individually aligned. Primers were designed for each consensus alignment for an amplicon size range of 90–300 bp. A maximum of five primer pairs were designed for each gRNA. An in silico PCR was performed against all target and non-target genomes with designed primers using isPcr to estimate the prevalence of each target amplicon across target and non-target genome databases. Similarly, all gRNAs were mapped to target and non-target genomes using BBMap to obtain an estimate of gRNA prevalence.

Data visualization

For visualization of gRNAs on bacterial genomes, gRNA sequences were mapped to one representative genome from each species using BLAST (v2.13.0)[79] with the flags "-qcov_hsp_perc 100 -perc_identity 100 -outfmt 6 -task blastn-short". Schematics of guide locations on the genomes were generated using the karyoploteR (v1.26.0)[80] package in R. All other graphs were generated using the ggplot2 (v3.4.2)[81] package in R. To determine genomic locations of gRNAs within the genome, one genome representative was selected for each species and annotated using Prokka to obtain coordinates of protein-coding regions. Sequences of gRNAs were mapped against their respective representative genomes using BLAST to obtain their genomic coordinates. Overlaps between the gRNAs with the genic and intergenic regions within each genome were determined using bedtools (v2.30.0)[82].

In silico primer and guide RNA design

For in silico testing of the PathoGD pipeline on five pathogens using both pangenome and *k*-mer modules, non-overlapping genome assemblies at the complete, chromosome, and scaffold levels were downloaded from NCBI RefSeq and GenBank (accessed June 27, 2023) for the target database for *T. pallidum* ($n = 92$), *C. trachomatis* ($n = 171$) and *S. pyogenes* ($n = 765$) based on taxonomic annotations obtained from the Genome Taxonomy Database (GTDB) r214[83]. Genome assemblies at all levels were retained for *N. gonorrhoeae* ($n = 978$) due to high genetic diversity between subtypes. Genomes were limited to complete and chromosome levels for *S. aureus* ($n = 1{,}096$) due to overrepresentation in the database. For the non-target database, genome assemblies at all levels were retained for *N. gonorrhoeae* ($n = 2597$), *T. pallidum* ($n = 642$), *C. trachomatis* ($n = 268$) and *S. aureus* ($n = 4600$), except for one non-target species for *S. pyogenes* ($n = 7057$), where 1000 genomes were subsampled from the total for *S. pneumoniae, S. agalactiae*, and *S. suis*. Genomes of all other species belonging to the same genus as the target species were used as the non-target database, except for *T. pallidum* and *C. trachomatis* where GTDB taxonomic discrepancies with NCBI were observed at the genus level. For *T. pallidum*, all genomes belonging to "g__Treponema", "g__Treponema_B", "g__Treponema_C", "g__Treponema_D", and "g__Treponema_F" in GTDB r214 were used for populating the non-target database. For *C. trachomatis*, all genomes belonging to "g__Chlamydia" and "g__Chlamydophila" in GTDB r214 were used for populating the non-target database. Additionally, the human reference genome (NCBI RefSeq assembly GCF_000001405.40) was included as part of the non-target database for all in silico designs. Primers and guide RNAs were designed using the PathoGD pangenome and *k*-mer modules with the 'user_all_subsample' workflow. For each species, a common subset of 100 target genomes was used for primer and gRNA design in both modules. All designed primers and gRNAs for each species were validated in silico against their respective full target databases. The metadata for all genome assemblies used was obtained from NCBI.

Comparison to gRNA designs from PrimedSherlock

Genomic sequences of Dengue virus serotype 1 (DENV-1; $n = 2230$), Dengue virus serotype 2 (DENV-2; $n = 776$), Dengue virus serotype 3 (DENV-3; $n = 946$), Dengue virus serotype 4 (DENV-4; $n = 214$), West Nile virus (WNV; $n = 2618$) and Zika virus (ZIKV; $n = 557$), and sequences from their respective non-target databases (DENV-1; $n = 5555$, DENV-2; $n = 6206$, DENV-3; $n = 4014$, DENV-4; $n = 3580$, WNV; $n = 6957$, ZIKV; $n = 9164$) and gRNA sequences were obtained from PrimedSherlock[29]. Primers and gRNAs were designed for the six viruses using the *k*-mer module of PathoGD without genome subsampling, with a specified amplicon size range of 90–500 base pairs, a gRNA prevalence threshold of 50%, and with the cd-hit amplicon sequence identity clustering threshold set to 80%. For DENV-3, the "--mismatch 1" flag was used for GenomeTester4 in the *k*-mer comparison steps as no output was obtained for the default setting. Primers and gRNAs from PathoGD and PrimedRPA/PrimedSherlock[29,32] for each species were evaluated based on the prevalence and conservation of their binding sites across all target and off-target genomes. Specifically, primers were assessed based on their potential to amplify their respective genomic regions using isPcr with the "-minPerfect=5" and "-maxSize=500" options, and prevalence across target genomes calculated for exact matches, ≤2 mismatches, ≤5 mismatches and ≤8 mismatches across the forward and reverse primers. Guide RNAs were evaluated based on the presence of the gRNA binding sites along with an upstream TTTN PAM in their respective target and non-target genome databases using GenomeTester4. For target genomes, gRNA comparisons were made with

no mismatches ("--mismatch 0" flag) or one mismatch allowed ("--mismatch 1" flag). For non-target genomes, comparisons were made allowing one ("--mismatch 1" flag), two ("--mismatch 2" flag), or three ("--mismatch 3" flag) mismatches.

Experimental validation of primers and guide RNAs with RPA-CRISPR-Cas12a assay

Genomic DNA (gDNA) from the target organisms, *S. pyogenes* and *N. gonorrhoeae*, as well as organisms in the specificity panel, were amplified using RPA with the TwistAmp Basic kit (TwistDx) according to the manufacturer's instructions. The approximate starting amounts for each organism ranged from approximately 9500–68,000 copies/μL (Supplementary Table 2). For the limit of detection assay, starting amounts of 4000, 400, and 40 copies/μL of *S. pyogenes* and *N. gonorrhoeae* gDNA were used. The RPA pellet was first resuspended with rehydration buffer supplied in the kit containing 0.48 μM of forward primer, 0.48 μM of reverse primer, and nuclease-free water to a total volume of 41 μL. The resuspended volume was split equally across four reactions, and 1.0 μL of template DNA and 1.25 μL of magnesium acetate (140 mM stock) were added. The reactions were incubated at 37 °C for 30 min in a T100 Thermal Cycler (Bio-Rad).

LbCas12a trans-cleavage assays were performed similarly to a previously described assay (13). A total of 40 nM LbCas12a (Gene Universal) was preincubated with 50 nM gRNA and 2 μL RPA product in 1× Holmes Buffer 2 in a black 96-well assay plate for 10 min at 37 °C (Hybaid hybridization oven). The /56-FAM/TTTTTTTTT/3BHQ-1/ reporter molecule (IDT) was added to the reaction at a final concentration of 125 nM. Fluorescence was measured using the CLARIOstar Plus Microplate Reader (BMG LABTECH) every minute for 38 min (λ_{ex} = 490/8 nm; λ_{em} = 520/8 nm) at 37 °C. No-template controls (NTC) were included as negative controls in all assays and contained all reagents except for template DNA. All RPA and CRISPR–Cas12a assays were performed with at least three experimental repeats, except for the *S. pyogenes* limit of detection assay, where one RPA assay was performed for each concentration, and four replicates were generated for the CRISPR–Cas12a assay.

For the amplification-free assays, LbCas12a trans-cleavage reactions were performed as above, with the exception that 250 ng/μL of *N. gonorrhoeae* gDNA was added. For the initial assessment of 6 multi-copy gRNAs, the assay was run for 58 min with 3 experimental repeats. For the pooled gRNA evaluation, the assay was run for either 58 (n = 5) or 118 min (n = 1), with 6 experimental repeats for all gRNAs and 3 experimental repeats for *porA*. Due to the high starting concentration required, specificity testing was performed only for the *porA* gRNA and reactions containing 4 pooled gRNAs, using 250 ng/μL of *N. lactamica*, *N. meningitidis*, and *N. subflava*, *C. trachomatis*, *E. coli*, and *S. pyogenes* gDNA. In silico analysis of the gRNA sequences confirmed no hits were observed against the remaining organisms in the *N. gonorrhoeae* specificity panel in Supplementary Table 2.

Statistics and reproducibility

All statistical analyses were performed using R version 4.2.2. The fluorescence values for the limit of detection assays were transformed using min-max normalization. To assess the significance of samples with target gDNA in comparison to the no-target control (NTC), a one-tailed t-test was conducted using the ggpubr package (v0.6.0) with the 'alternative = "greater"' parameter.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The accession number of genomes used and/or analyzed during the current study is available in the Supplementary Data files. The genome assemblies used for the in silico design of primers and guide RNAs for the five pathogens are available in Supplementary Data 1. The source data used for generating the main figures are available in Supplementary Data 4. Sequences and metadata for the primers and guide RNAs designed in silico are available from the corresponding author upon request.

Received: 23 July 2024; Accepted: 22 January 2025;
Published online: 30 January 2025

References

1. Mahony, J. B. et al. Cost analysis of multiplex PCR testing for diagnosing respiratory virus infections. *J. Clin. Microbiol.* **47**, 2812–2817 (2009).
2. Land, K. J., Boeras, D. I., Chen, X. S., Ramsay, A. R. & Peeling, R. W. REASSURED diagnostics to inform disease control strategies, strengthen health systems and improve patient outcomes. *Nat. Microbiol.* **4**, 46–54 (2019).
3. Kaminski, M. M., Abudayyeh, O. O., Gootenberg, J. S., Zhang, F. & Collins, J. J. CRISPR-based diagnostics. *Nat. Biomed. Eng.* **5**, 643–656 (2021).
4. Gootenberg, J. S. et al. Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science* **356**, 438–442 (2017).
5. Myhrvold, C. et al. Field-deployable viral diagnostics using CRISPR-Cas13. *Science* **360**, 444–448 (2018).
6. Chen, J. S. et al. CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science* **360**, 436–439 (2018).
7. Ai, J. W. et al. CRISPR-based rapid and ultra-sensitive diagnostic test for Mycobacterium tuberculosis. *Emerg. Microbes Infect.* **8**, 1361–1369 (2019).
8. Cunningham, C. H. et al. A novel CRISPR-based malaria diagnostic capable of Plasmodium detection, species differentiation, and drug-resistance genotyping. *EBioMedicine* **68**, 103415 (2021).
9. Lee, R. A. et al. Ultrasensitive CRISPR-based diagnostic for field-applicable detection of Plasmodium species in symptomatic and asymptomatic malaria. *Proc. Natl Acad. Sci. USA* **117**, 25722–25731 (2020).
10. Broughton, J. P. et al. CRISPR-Cas12-based detection of SARS-CoV-2. *Nat. Biotechnol.* **38**, 870–874 (2020).
11. Patchsung, M. et al. Clinical validation of a Cas13-based assay for the detection of SARS-CoV-2 RNA. *Nat. Biomed. Eng.* **4**, 1140–1149 (2020).
12. Welch, N. L. et al. Multiplexed CRISPR-based microfluidic platform for clinical testing of respiratory viruses and identification of SARS-CoV-2 variants. *Nat. Med.* **28**, 1083–1094 (2022).
13. Low, S. J. et al. Rapid detection of monkeypox virus using a CRISPR-Cas12a mediated assay: a laboratory validation and evaluation study. *Lancet Microbe* **4**, E800–E810 (2023).
14. Chen, Q. et al. CRISPR-Cas12-based field-deployable system for rapid detection of synthetic DNA sequence of the monkeypox virus genome. *J. Med. Virol.* **95**, e28385 (2023).
15. Kellner, M. J., Koob, J. G., Gootenberg, J. S., Abudayyeh, O. O. & Zhang, F. SHERLOCK: nucleic acid detection with CRISPR nucleases. *Nat. Protoc.* **14**, 2986–3012 (2019).
16. Abudayyeh, O. O. et al. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* **353**, aaf5573 (2016).
17. East-Seletsky, A. et al. Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature* **538**, 270–273 (2016).
18. Tu, Q. et al. Point-of-care detection of Neisseria gonorrhoeae based on RPA-CRISPR/Cas12a. *AMB Express* **13**, 50 (2023).
19. Liu, L. et al. Generation and application of a novel high-throughput detection based on RPA-CRISPR technique to sensitively monitor pathogenic microorganisms in the environment. *Sci. Total Environ.* **838**, 156048 (2022).
20. Hicks, A. L. et al. Targeted surveillance strategies for efficient detection of novel antibiotic resistance variants. *Elife* **9**, e56367 (2020).

21. Fonfara, I., Richter, H., Bratovic, M., Le Rhun, A. & Charpentier, E. The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature* **532**, 517–521 (2016).
22. Chuai, G. et al. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biol.* **19**, 80 (2018).
23. Poudel, R., Rodriguez, L. T., Reisch, C. R. & Rivers, A. R. GuideMaker: software to design CRISPR-Cas guide RNA pools in non-model genomes. *Gigascience* **11**, giac007 (2022).
24. Moreno-Mateos, M. A. et al. CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat. Methods* **12**, 982–988 (2015).
25. Zhu, H. & Liang, C. CRISPR-DT: designing gRNAs for the CRISPR-Cpf1 system with improved target efficiency and specificity. *Bioinformatics* **35**, 2783–2789 (2019).
26. Zhu, L. J., Holmes, B. R., Aronin, N. & Brodsky, M. H. CRISPRseek: a bioconductor package to identify target-specific guide RNAs for CRISPR-Cas9 genome-editing systems. *PLoS ONE* **9**, e108424 (2014).
27. Perez, A. R. et al. GuideScan software for improved single and paired CRISPR guide RNA design. *Nat. Biotechnol.* **35**, 347–349 (2017).
28. Hoberecht, L., Perampalam, P., Lun, A. & Fortin, J. P. A comprehensive Bioconductor ecosystem for the design of CRISPR guide RNAs across nucleases and technologies. *Nat. Commun.* **13**, 6568 (2022).
29. Mann, J. G. & Pitts, R. J. PrimedSherlock: a tool for rapid design of highly specific CRISPR-Cas12 crRNAs. *BMC Bioinforma.* **23**, 428 (2022).
30. Ansari, A. H., Kumar, M., Sarkar, S., Maiti, S. & Chakraborty, D. CriSNPr, a single interface for the curated and de novo design of gRNAs for CRISPR diagnostics using diverse Cas systems. *Elife* **12**, e77976 (2023).
31. Foster, Z. S. L., Tupper, A. S., Press, C. M. & Grunwald, N. J. Krisp: A Python package to aid in the design of CRISPR and amplification-based diagnostic assays from whole genome sequencing data. *PLoS Comput. Biol.* **20**, e1012139 (2024).
32. Higgins, M. et al. PrimedRPA: primer design for recombinase polymerase amplification assays. *Bioinformatics* **35**, 682–684 (2019).
33. Kitts, P. A. et al. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.* **44**, D73–D80 (2016).
34. Kirchberger, P. C., Schmidt, M. L. & Ochman, H. The Ingenuity of Bacterial Genomes. *Annu. Rev. Microbiol.* **74**, 815–834 (2020).
35. Land, M. et al. Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics* **15**, 141–161 (2015).
36. Thorpe, H. A., Bayliss, S. C., Hurst, L. D. & Feil, E. J. Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. *Genetics* **206**, 363–376 (2017).
37. Marchais, A., Naville, M., Bohn, C., Bouloc, P. & Gautheret, D. Single-pass classification of all noncoding sequences in a bacterial genome using phylogenetic profiles. *Genome Res.* **19**, 1084–1092 (2009).
38. Daher, R. K., Stewart, G., Boissinot, M., Boudreau, D. K. & Bergeron, M. G. Influence of sequence mismatches on the specificity of recombinase polymerase amplification technology. *Mol. Cell Probes* **29**, 116–121 (2015).
39. Munawar, M. A. Critical insight into recombinase polymerase amplification technology. *Expert Rev. Mol. Diagn.* **22**, 725–737 (2022).
40. Boyle, D. S. et al. Rapid detection of HIV-1 proviral DNA for early infant diagnosis using recombinase polymerase amplification. *mBio* **4**, e00135–13 (2013).
41. Strohkendl, I., Saifuddin, F. A., Rybarski, J. R., Finkelstein, I. J. & Russell, R. Kinetic basis for DNA target specificity of CRISPR-Cas12a. *Mol. Cell* **71**, 816–824 e813 (2018).
42. Li, S. Y. et al. CRISPR-Cas12a-assisted nucleic acid detection. *Cell Discov.* **4**, 20 (2018).
43. Swarts, D. C., van der Oost, J. & Jinek, M. Structural basis for guide RNA processing and seed-dependent DNA targeting by CRISPR-Cas12a. *Mol. Cell* **66**, 221–233 e224 (2017).
44. Yamano, T. et al. Structural basis for the canonical and non-canonical PAM recognition by CRISPR-Cpf1. *Mol. Cell* **67**, 633–645 e633 (2017).
45. Toth, E. et al. Improved LbCas12a variants with altered PAM specificities further broaden the genome targeting range of Cas12a nucleases. *Nucleic Acids Res.* **48**, 3722–3733 (2020).
46. World Health Organization, Vol. 2023 (World Health Organization, 2016).
47. Vekemans, J. et al. The path to Group A streptococcus vaccines: World Health Organization Research and Development Technology Roadmap and preferred product characteristics. *Clin. Infect. Dis.* **69**, 877–883 (2019).
48. World Health Organization, Vol. 2023 (World Health Organization, 2017).
49. Kaplinski, L., Lepamets, M. & Remm, M. GenomeTester4: a toolkit for performing basic set operations - union, intersection and complement on k-mer lists. *Gigascience* **4**, 58 (2015).
50. Untergasser, A. et al. Primer3-new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012).
51. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144–161 (2013).
52. Pla-Diaz, M. et al. Evolutionary processes in the emergence and recent spread of the syphilis agent, *Treponema pallidum*. *Mol. Biol. Evol.* **39**, msab318 (2022).
53. Abdelsamed, H., Peters, J. & Byrne, G. I. Genetic variation in *Chlamydia trachomatis* and their hosts: impact on disease severity and tissue tropism. *Future Microbiol* **8**, 1129–1146 (2013).
54. Pinto, M. et al. Genome-scale analysis of the non-cultivable *Treponema pallidum* reveals extensive within-patient genetic variation. *Nat. Microbiol* **2**, 16190 (2016).
55. Fyfe, J. A. et al. Development and application of two multiplex real-time PCR assays for the detection of *Mycobacterium ulcerans* in clinical and environmental samples. *Appl. Environ. Microbiol.* **73**, 4733–4740 (2007).
56. Register, K. B. & Sanden, G. N. Prevalence and sequence variants of IS481 in *Bordetella bronchiseptica*: implications for IS481-based detection of *Bordetella pertussis*. *J. Clin. Microbiol.* **44**, 4577–4583 (2006).
57. Dong, Q. et al. Ultrasensitive detection of pathogenic bacteria by targeting high copy signature genes. *Front. Vet. Sci.* **9**, 889419 (2022).
58. Fozouni, P. et al. Amplification-free detection of SARS-CoV-2 with CRISPR-Cas13a and mobile phone microscopy. *Cell* **184**, 323–333 e329 (2021).
59. Marjuki, H. et al. Genetic similarity of gonococcal homologs to meningococcal outer membrane proteins of serogroup B vaccine. *mBio* **10**, e01668–19 (2019).
60. Sparling, P. F. Genetic transformation of Neisseria gonorrhoeae to streptomycin resistance. *J. Bacteriol.* **92**, 1364–1371 (1966).
61. Spratt, B. G., Bowler, L. D., Zhang, Q. Y., Zhou, J. & Smith, J. M. Role of interspecies transfer of chromosomal genes in the evolution of penicillin resistance in pathogenic and commensal Neisseria species. *J. Mol. Evol.* **34**, 115–125 (1992).
62. Cehovin, A. & Lewis, S. B. Mobile genetic elements in Neisseria gonorrhoeae: movement for change. *Pathog. Dis.* **75**, ftx071 (2017).
63. Whiley, D. M., Tapsall, J. W. & Sloots, T. P. Nucleic acid amplification testing for Neisseria gonorrhoeae: an ongoing challenge. *J. Mol. Diagn.* **8**, 3–15 (2006).
64. Tabrizi, S. N. et al. Evaluation of six commercial nucleic acid amplification tests for detection of Neisseria gonorrhoeae and other Neisseria species. *J. Clin. Microbiol.* **49**, 3610–3615 (2011).
65. Luo, H. et al. Development and application of Cas13a-based diagnostic assay for Neisseria gonorrhoeae detection and azithromycin resistance identification. *J. Antimicrob. Chemother.* **77**, 656–664 (2022).

66. Allan-Blitz, L. T. et al. Development of Cas13a-based assays for Neisseria gonorrhoeae detection and gyrase A determination. *mSphere* **8**, e0041623 (2023).
67. Nalefski, E. A. et al. Kinetic analysis of Cas12a and Cas13a RNA-Guided nucleases for development of improved CRISPR-Based diagnostics. *iScience* **24**, 102996 (2021).
68. Li, H. et al. Amplification-free CRISPR/Cas detection technology: challenges, strategies, and perspectives. *Chem. Soc. Rev.* **52**, 361–382 (2023).
69. Zhang, J. et al. Recent Improvements in CRISPR-Based Amplification-Free Pathogen Detection. *Front. Microbiol.* **12**, 751408 (2021).
70. Liu, G., Zhang, Y. & Zhang, T. Computational approaches for effective CRISPR guide RNA design and evaluation. *Comput. Struct. Biotechnol. J.* **18**, 35–44 (2020).
71. Metsky, H. C. et al. Designing sensitive viral diagnostics with machine learning. *Nat. Biotechnol.* **40**, 1123–1131 (2022).
72. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
73. Page, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
74. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinforma.* **11**, 119 (2010).
75. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
76. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
77. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS ONE* **11**, e0163962 (2016).
78. Bushnell, B. BBMap short-read aligner, and other bioinformatics tools. https://sourceforge.net/projects/bbmap/ (2016).
79. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
80. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).
81. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).
82. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
83. Parks, D. H. et al. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).

Acknowledgements

This study was funded by the Victorian Medical Research Accelerator Fund (GA-F3791196–5514) and the Australian Government Department of Health (PO4932). D.A.W. is supported by an NHMRC Investigator Grant (APP1174555). This work was also supported by an Australian Research Council Industrial Transformation Research Hub Grant (IH190100021). This work was performed using computational resources at the Walter and Eliza Hall Institute of Medical Research.

Author contributions

D.A.W., S.P., and S.J.L. designed the study. D.A.W. and S.P. provided supervision and support. S.J.L. developed the pipeline and performed bioinformatic analyses, with input from G.T., L.W. and S.P. M.O., W.J.K., and N. W. performed the experimental validations, with contributions from M.K., F.A, Y.N., and E.H. E.S. provided expertise in software development. S.J.L., S.P., and D.A.W. wrote the initial draft of the paper. All authors read and approved the final paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-025-07591-1.

Correspondence and requests for materials should be addressed to Soo Jen Low or Shivani Pasricha.

Peer review information *Communications Biology* thanks Zachary S. L. Foster and the other, anonymous, reviewers for their contribution to the peer review of this work. Primary Handling Editor: Mengtan Xing. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025