scientific data

OPEN

ARTICLE

CF-Seq, an accessible web application for rapid re-analysis of cystic fibrosis pathogen RNA sequencing studies

Samuel L. Neff[1], Thomas H. Hampton[1], Charles Puerner[1], Liviu Cengher[1], Georgia Doing[1], Alexandra J. Lee[2], Katja Koeppen[1], Ambrose L. Cheung[1], Deborah A. Hogan[1], Robert A. Cramer[1] & Bruce A. Stanton[1] ✉

Researchers studying cystic fibrosis (CF) pathogens have produced numerous RNA-seq datasets which are available in the gene expression omnibus (GEO). Although these studies are publicly available, substantial computational expertise and manual effort are required to compare similar studies, visualize gene expression patterns within studies, and use published data to generate new experimental hypotheses. Furthermore, it is difficult to filter available studies by domain-relevant attributes such as strain, treatment, or media, or for a researcher to assess how a specific gene responds to various experimental conditions across studies. To reduce these barriers to data re-analysis, we have developed an R Shiny application called CF-Seq, which works with a compendium of 128 studies and 1,322 individual samples from 13 clinically relevant CF pathogens. The application allows users to filter studies by experimental factors and to view complex differential gene expression analyses at the click of a button. Here we present a series of use cases that demonstrate the application is a useful and efficient tool for new hypothesis generation. (CF-Seq: http://scangeo.dartmouth.edu/CFSeq/)

Introduction

Cystic fibrosis (CF) is a monogenic, homozygous recessive genetic disease that affects over 30,000 people in the US and more than 70,000 worldwide[1]. The disease is caused by mutations of the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene, which is expressed in a wide variety of cells throughout the body but has been predominantly studied in the context of the lungs and the digestive system[2–5]. In the lungs, the absence of CFTR protein contributes to mucus obstruction, chronic microbial infections, systemic inflammation, and progressive lung disease, which is the leading cause of mortality[6–9]. Furthermore, people with CF (pwCF) are commonly diagnosed with exocrine pancreatic insufficiency, and tend to exhibit microbial dysbiosis in the GI tract, which both contribute to nutritional deficits, poor growth, and a myriad of other GI symptoms[5,10,11].

Based on population data from the Cystic Fibrosis Foundation Patient Registry, pwCF born between 2015 and 2019 have a median life expectancy of 46 years[12]. Further improvements in life expectancy are imminent, given the advent of highly effective CF modulator therapies (HEMT) over the past decade[13]. CF modulators were first made accessible with the approval of Ivacaftor in 2012 for a small subset of CF patients and access was expanded to most CF patients with the approval of Trikafta in 2019. These drugs have brought improvements in lung function and nutritional status, while decreasing the frequency of hospitalization for pulmonary exacerbation[14]. In recent studies, CF modulator treatment has also been associated with a reduced incidence of infection with various common CF lung pathogens[14–17]. There is hope that young patients with stable lung disease may avoid chronic infection with these pathogens altogether, though studies conducted so far on this subject suggest that the modulator drugs may not be capable of eradicating established infection in older patients[18]. Thus, further development of maintenance and/or eradication therapy to treat infection is still a necessity.

Given the contribution of invasive pathogens to lung disease progression, lung microbiology has long been a key focus of CF research. CF researchers have traditionally studied a suite of "classic CF pathogens" that are

[1]Geisel School of Medicine, Dartmouth College, Hanover, NH, USA. [2]University of Pennsylvania, Philadelphia, PA, USA. ✉e-mail: Bruce.A.Stanton@dartmouth.edu

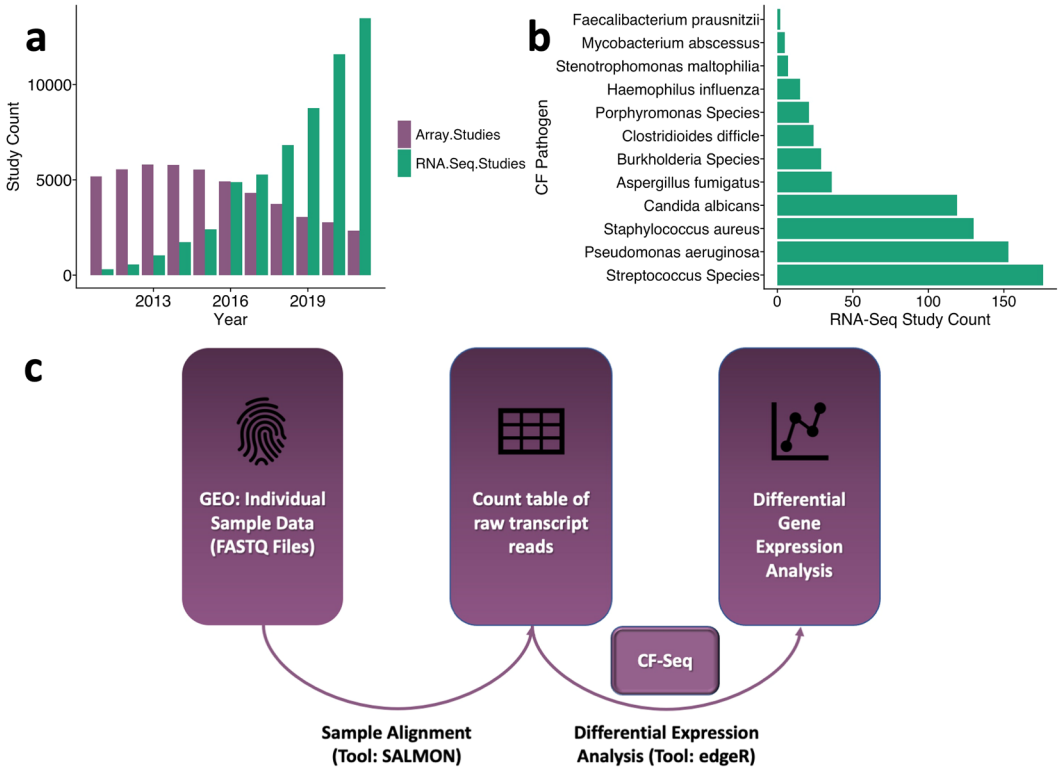

Fig. 1 Landscape of RNA-Sequencing studies available in the Gene Expression Omnibus (GEO). (**a**) Since 2011, the number of RNA-Seq studies hosted in GEO has risen dramatically, from several hundred to over 10,000, well eclipsing the number of microarray expression studies currently produced per year. (**b**) While small relative to the total set of RNA-Seq studies in GEO, there is a substantial number of RNA-Seq studies available for the CF pathogen species featured in the CF-Seq application. (**c**) To derive meaningful biological insights from the RNA-seq studies in GEO, the analysis pipeline outlined here must be followed. Alignment of sample RNA sequences to a reference genome is accomplished with a command line tool like SALMON, and downstream analysis with a tool such as the popular R package edgeR. CF-Seq automates the second segment of this pipeline, saving users from the need to clean up count tables, produce experimental design matrices, gather metadata, and write sophisticated analysis code in R.

known to infect the CF lungs and exacerbate lung disease. These pathogens include the gram-negative bacterium *Pseudomonas aeruginosa*, the gram-positive bacterium *Staphylococcus aureus*, gram-negative bacteria of the genus *Burkholderia*, and fungal species such as *Aspergillus fumigatus* [Supplemental Table S1]. In recent years, the set of recognized CF pathogens has expanded as epidemiological studies have identified species that are rising in prevalence and impacting clinical outcomes (e.g., non-tuberculous mycobacteria species such as *M. abscessus*)[19,20]. In addition, more sensitive culture tools have allowed researchers to recognize the clinical relevance of less prevalent aerobic and anaerobic species[21,22]. Recently, researchers have begun to develop model systems to interrogate the interactions between CF pathogens in the lungs and to consider how the overall shape of the CF community – the diversity and abundance of different bacteria – contributes to clinical outcomes[23]. In fact, studies have found that a patient's microbial community as a whole may be more effective at predicting disease outcomes than colonization with any individual species[24].

Decades of prior CF pathogen research has helped advise modern clinical treatments, and this published body of research continues to serve as a source of knowledge for drug development as well as inspiration for future studies. High-throughput transcriptomics experiments – of which RNA-Seq studies have recently become most common – are especially useful as a source of published data to inform future experiments Many such datasets are presently available in the NCBI Gene Expression Omnibus (GEO) [Fig. 1a,b]. In an ideal world, CF pathogen researchers would be able to view which microbial strains, treatment conditions, and media have previously been utilized, and perform a quick visual analysis of gene expression under these conditions. This information would offer researchers a roadmap to identify future directions for follow-up experiments. However, we do not (yet) live in this ideal world. Although many datasets are publicly available, substantial computational expertise and manual effort are required to compare similar studies, visualize gene expression patterns within studies, and use published data to generate new experimental hypotheses. Thus, there is a need develop an application that will reduce these barriers to data re-analysis.

One useful approach to derive biological insights from an RNA-Seq dataset in GEO – and the one that we automate in the CF-Seq application – is to see which genes are differentially expressed under varying experimental conditions. To accomplish this analysis, a researcher would first need to locate the sample runs associated with the individual dataset. These are often stored as FASTQ files that require extensive computational skills

to process. Someone with these skills could trim the sequence reads contained in the FASTQ files to remove low quality reads and adapter sequences. Next, they would align trimmed reads to a reference genome with a command-line tool like SALMON[25], which yields a count table with raw gene expression counts for each sample. Then, finally, that researcher could conduct differential gene expression analysis. This final step requires knowledge of a programming language like R[26,27], and specific R packages like edgeR[28,29] or DESeq[30] that allow for the generation of biologically meaningful analysis tables and figures. Even among bioinformatics researchers, many do not have expertise in all aspects of this pipeline – and for those who do, running through the pipeline for just a single dataset is typically a multi-day effort. CF-Seq has been designed so that users do not have to deal with this pipeline at all. Taking advantage of count tables that dataset contributors have left in GEO as supplemental files, CF-Seq takes care of differential expression analysis [Fig. 1c].

Our efforts to make public data more accessible are certainly not the first of their kind. In recent years, as big -omics datasets have become increasingly commonplace and researchers have encountered the challenges described above, the necessity of adopting FAIR data principles by making datasets more Findable, Accessible, Interoperable, and Reproducible has increasingly been recognized[31]. In this spirit, various research tools have already been developed to make publicly available data more amenable to re-use. For example, the application *MetaRNA-Seq* enables users to view consolidated study metadata that had been scattered across the four NCBI databases: SRA, Biosample, Bioprojects, and GEO[32]. Another application, the *geoCancerPrognosticDatasets Retriever*, allows users to utilize additional search parameters (e.g., cancer type) to retrieve GEO accessions for all studies of interests[33].

Some existing applications designed by other research teams are quite similar in nature to CF-Seq and have served as strong inspiration for our own efforts. However, none are specifically geared towards CF pathogen research, and there is room to expand on their functionality [Supplemental Table S2]. Our own lab has previously published tools to make publicly available data more accessible to CF researchers[34,35], but these tools focus on the most studied CF pathogens – namely *Pseudomonas aeruginosa* and *Staphylococcus aureus* – and don't include datasets on many of the other clinically relevant species listed in Supplemental Table S1.

Building on our prior work, we present the R Shiny web application CF-Seq. CF-Seq is a web application based on a compendium of RNA-Seq experiments. This compendium contains 13 clinically relevant CF pathogens; a mix of aerobes and anaerobes residing in the lung and the digestive tract. The application currently holds carefully formatted count tables and metadata for 128 studies, and 1,322 RNA-seq samples in total, with ongoing efforts to capture more studies and additional relevant species, as outlined in the Discussion section. All datasets currently included in the application are arranged by GEO accession number in Supplemental Table S3 for reference.

The CF-Seq application allows differential gene expression analysis of each individual study at the click of a button, producing downloadable tables and figures depicting fold changes and p values of differentially expressed genes in a matter of seconds. For each study, the application allows users to produce tables and figures comparing individual sample groups (e.g., samples treated with antibiotic X vs. control samples, samples treated with antibiotic Y vs. samples treated with antibiotic X, etc.). For many species and strains (where KEGG pathway annotations are available) the user can also visualize how the genes in specific biological pathways are differentially expressed. Furthermore, the user can filter all studies on the same species – breaking them down by strain, media, treatment, or gene(s) perturbed – to identify all past experimental conditions (and combinations of conditions) and thus determine which have yet to be assessed [Fig. 2]. This application has been developed with the close guidance of CF pathogen researchers at the Geisel School of Medicine at Dartmouth College. In this publication, we present three case studies that showcase the application's usefulness for researchers studying three different CF pathogens (*Aspergillus fumigatus*, *Pseudomonas aeruginosa*, and *Staphylococcus aureus*).

Results

The CF-Seq application makes it simple for CF researchers to take full advantage of the 128 CF RNA-Seq pathogen datasets in the associated compendium. Upon opening the application, the user is greeted with a user manual that instructs them on how best to use CF-Seq [Fig. 2, Panel 1]. After reading, the user is then directed to the central, study-filtering panel of the application [Fig. 2, Panel 2]. Here, the user can filter studies by species, and then by strain, media, treatment, or gene perturbation [Fig. 2, Panel 3]. Filtered studies are presented in a table and can be selected to reveal additional metadata – including the study name, description, and link to its record in GEO [Fig. 2, Panel 4]. Once a study is selected, the user can click a button to reveal detailed differential expression analysis in a separate analysis tab [Fig. 2, Panel 5]. This analysis includes a table with the fold change (FC), p value, and counts per million (CPM) of all genes assessed in the study. For species or strains in which KEGG pathway information is available, the user is also able to visualize how the genes on different KEGG pathways are up or downregulated [Fig. 2, Panel 6]. Certain studies also allow users to visualize genes associated with various Gene Ontology (GO) terms, Clusters of Orthologous Genes (COG) categories, or functional descriptions of gene activity.

A series of user stories have been developed by three of the publication co-authors to demonstrate the value of the application in a research setting. These co-authors conduct research in laboratories that frequently publish papers related to CF microbiology. The following section of the manuscript demonstrates the analysis features of the application and outlines how these researchers used the application to come up with new questions and testable hypotheses relevant to their own research. Given the current focus in the field of CF research on the CF microbiome as a polymicrobial community[23,24], all three user stories focus on polymicrobial interactions between several CF pathogens. All volcano plots used as figures for the user stories were taken directly from the application.

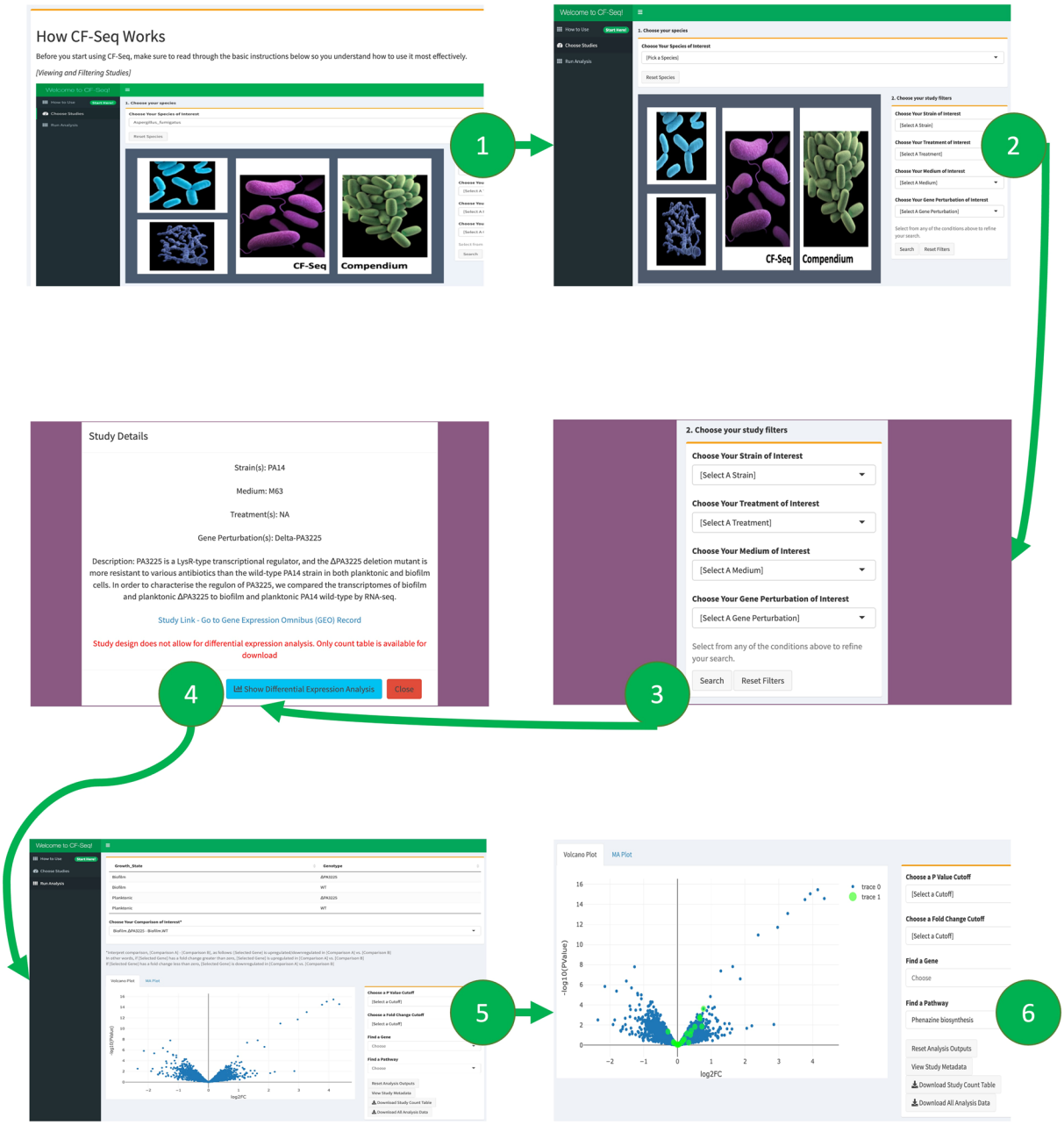

Fig. 2 Application workflow for CF-Seq users. Panel 1 shows the starting window of the application, where users are presented with a manual that explains the functionality and purpose of the application. Users are then directed to the study view screen, shown in panel 2, where they can select a species of interest and view available RNA-Seq studies. Panel 3 shows how filters can be applied to delineate studies with certain experimental characteristics (strain, media, treatment, gene perturbed). Panel 4 offers a look at the metadata that can be examined for each individual study. Panels 5 and 6 show the study analysis window, where analysis tables and figures can be generated for all experimental comparisons, individual genes may be highlighted, P value and fold change cutoffs can be selected, and differentially expressed genes on selected KEGG pathways can be highlighted when KEGG pathway information is available (Panel 6). For certain studies, users can also highlight other biological features, such as GO terms, COG categories, and functional descriptions of genes (e.g., "serine/threonine protein kinase") Zoomed-in versions of the figure panels showing more detail are available as Supplementary Figures S1–S6.

Case study #1: Examining aspergillus fumigatus in bacterial co-culture. *Dr. Charles Puerner, Cramer Laboratory, Geisel School of Medicine.*

The infectious mold *Aspergillus fumigatus* is ubiquitous in the environment[36]. The spores from this fungus are taken into the lung by breathing and normally cleared by a healthy immune system. However, individuals with compromised immune systems and pulmonary diseases such as cystic fibrosis are particularly vulnerable

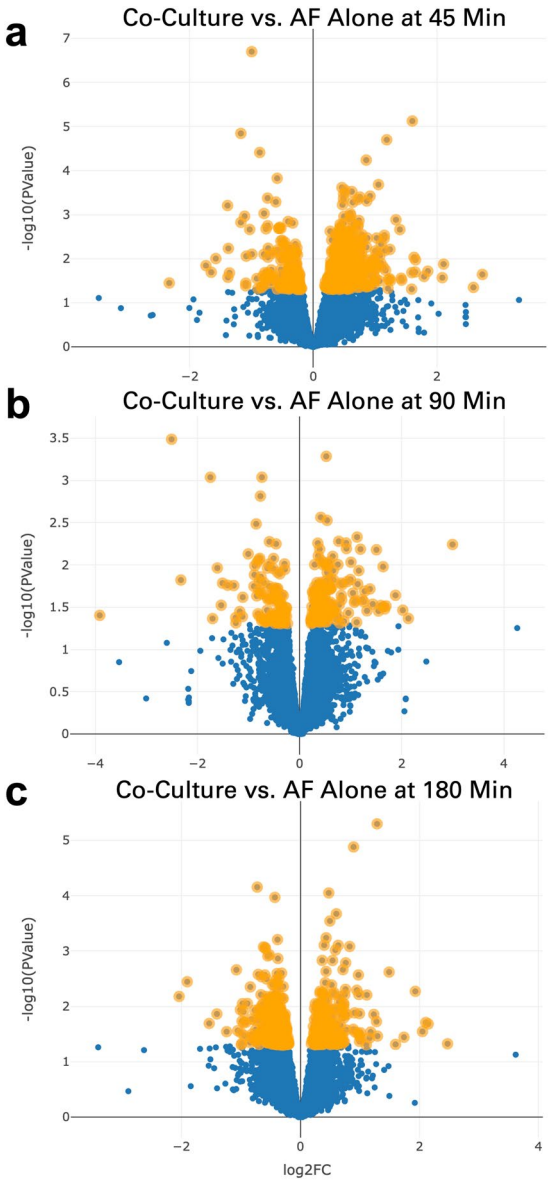

Fig. 3 Expression of A. fumigatus genes following exposure to P. aeruginosa presented in volcano plot format. In the CF-Seq application, the species A. fumigatus strain A1160 was selected and the dataset "Transcriptomics analysis of Aspergillus fumigatus co-cultivated with Pseudomonas aeruginosa" (GSE122391) was used for the subsequent analysis. Comparisons were selected comparing fungus co-cultured with bacteria to fungus alone at (**a**) 45, (**b**) 90, and (**c**) 180 min. Genes highlighted in orange are those whose p-value was less than 0.05. At 45 minutes, 531 of 8526 total genes were differentially expressed to a statistically significant degree. At 90 minutes and 180 minutes, the number of statistically significant differentially expressed genes was 257 and 514 respectively.

to infection by this fungus. In these cases, *A. fumigatus* spores are capable of germinating in the lung environment and forming fungal lesions. The Cramer lab studies the biology of this organism, specifically as it relates to its disease-causing capabilities. A recent publication, for example, investigated the genetic characteristics of persistent isolates taken from the lungs of a CF patient over several years[37].

Using the analysis capabilities of this application, we were particularly interested in a dataset which generated gene expression profiles of *A. fumigatus* co-cultured with the ubiquitous bacterium *Pseudomonas aeruginosa* (GEO: GSE122391). This dataset is interesting because both organisms are commonly found in the CF lung environment, a situation associated with worsened disease state[38]. The study was identified using the CF-Seq filtering feature to focus on those experiments that involved cross-species interactions.

In the analysis window of the application, the "Choose a P Value Cutoff" field was used to highlight genes whose p-value was 0.05 or less. Genes were highlighted at several timepoints comparing the co-culture of *P. aeruginosa* with *A. fumigatus* to culture of the fungus alone [Fig. 3]. The volcano and MA plots demonstrating the magnitude of differential expression, as well as a spreadsheet of statistically significant differentially

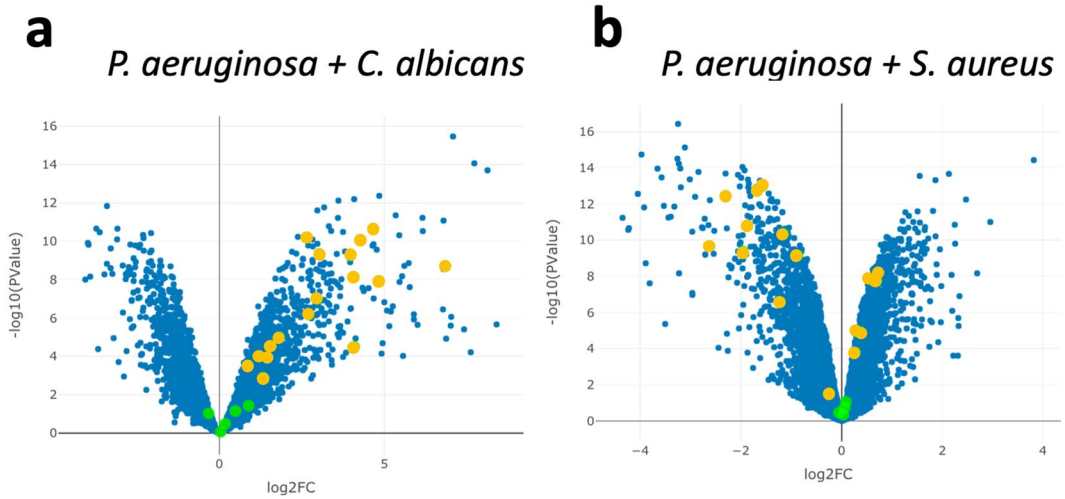

Fig. 4 (**a**) P. aeruginosa genes involved in phenazine biosynthesis tend to be upregulated in co-culture with C. albicans (**b**) but not in co-culture with S. aureus compared to P. aeruginosa in monoculture. Green data points were highlighted by selecting the KEGG pathway for phenazine biosynthesis using the 'find a pathway' feature in the CF-Seq application. Genes that are differentially expressed between co-culture and monoculture conditions to a statistically significant degree ($p < 0.05$) were colored orange for emphasis.

expressed genes, were quickly downloaded for further analysis and additional figure generation. The downloaded table of differentially expressed genes was easily filtered outside of the application to contain only genes with a $|\log_2 FC|$ value of 1.5 or greater (Fold change $> 2.83 = \log2FC > 1.5$, Fold change $< 0.35 = \log2FC < -1.5$) [Supplemental Table S4].

Using this method, the application makes it easy to identify a list of biologically significant genes which could be investigated further regarding their role in the co-culture environment. Genes differentially expressed with an especially high fold change and p value may be manipulated in the laboratory to see how the knockout of individual genes effects survival fitness of *A. fumigatus* in co-culture.

Case study #2: P. aeruginosa virulence factor production in polymicrobial contexts. Dr. Georgia Doing, Hogan Laboratory, Geisel School of Medicine.

Pseudomonas aeruginosa is one of the most common pathogens associated with cystic fibrosis (CF) lung infections, remains difficult to treat with antibiotics, and is associated with lung function decline in colonized pwCF[39]. Along with its ability to form recalcitrant biofilms and resist antibiotic treatment, its behaviors during interactions with other bacteria are now recognized as important factors that influence *P. aeruginosa* infection outcomes[40–44]. Microbial interactions are often studied in the laboratory using co-cultures of *P. aeruginosa* with other CF pathogens such as *Candida albicans* and *Staphylococcus aureus*. These co-culture experiments have proven to be useful for modeling polymicrobial interactions. However, it is increasingly apparent that the combinatorial effects of environmental factors as well as pairwise and community-wide microbial interactions make for complex systems with many changing variables and a large search space[44–46]. In addition to conducting new experiments in the laboratory, the re-analysis of individual datasets related to bacterial co-culture and meta-analysis of multiple datasets will likely spur new experimental hypotheses and help provide evidence for existing theories of polymicrobial interactions.

Using CF-Seq it was easy to compare two datasets from experiments where *P. aeruginosa* was co-cultured with *C. albicans* (GEO: GSE148597)[45] and (GEO: GSE122048) *S. aureus*[47]. We noticed that while *P. aeruginosa* mainly upregulates and highly expresses genes in the KEGG pathway for phenazine biosynthesis in co-cultures with *C. albicans* compared to monoculture [Fig. 4a], it does not do so in co-culture with *S. aureus* compared to monoculture [Fig. 4b]. Since *P. aeruginosa* phenazine production is induced with *C. albicans* fermentation[48], we searched for specific genes whose expression could indicate differences in *C. albicans* and *S. aureus* metabolisms that may shed light on their different effects on *P. aeruginosa* phenazine production.

Digging deeper into the data on an individual gene level, the upregulation of lactate permeases and lactate dehydrogenases by *P. aeruginosa* in co-culture with either *C. albicans* or *S. aureus* suggest both *C. albicans* and *S. aureus* were producing lactate in these experiments [Fig. 5a]. However, while *P. aeruginosa* upregulated alcohol dehydrogenases in co-culture with *C. albicans*, it did not do so in co-culture with *S. aureus*, suggesting *C. albicans* was likely producing ethanol while *S. aureus* was not [Fig. 5b]. Amongst the many differences between these two co-cultures, differences in microbially-produced fermentation products could lead to differences in *P. aeruginosa* phenazine production.

P. aeruginosa regulates phenazine production in response to quorum sensing and many metabolic and stress-related cues. Since both co-culture with *C. albicans* and co-culture with *S. aureus* elicited lactate metabolism, but only co-culture with *C. albicans* elicited ethanol metabolism, CF-Seq analysis suggests that ethanol specifically promotes phenazine production while lactate does not [Fig. 5c]. Thus, lactate may have a neutral or

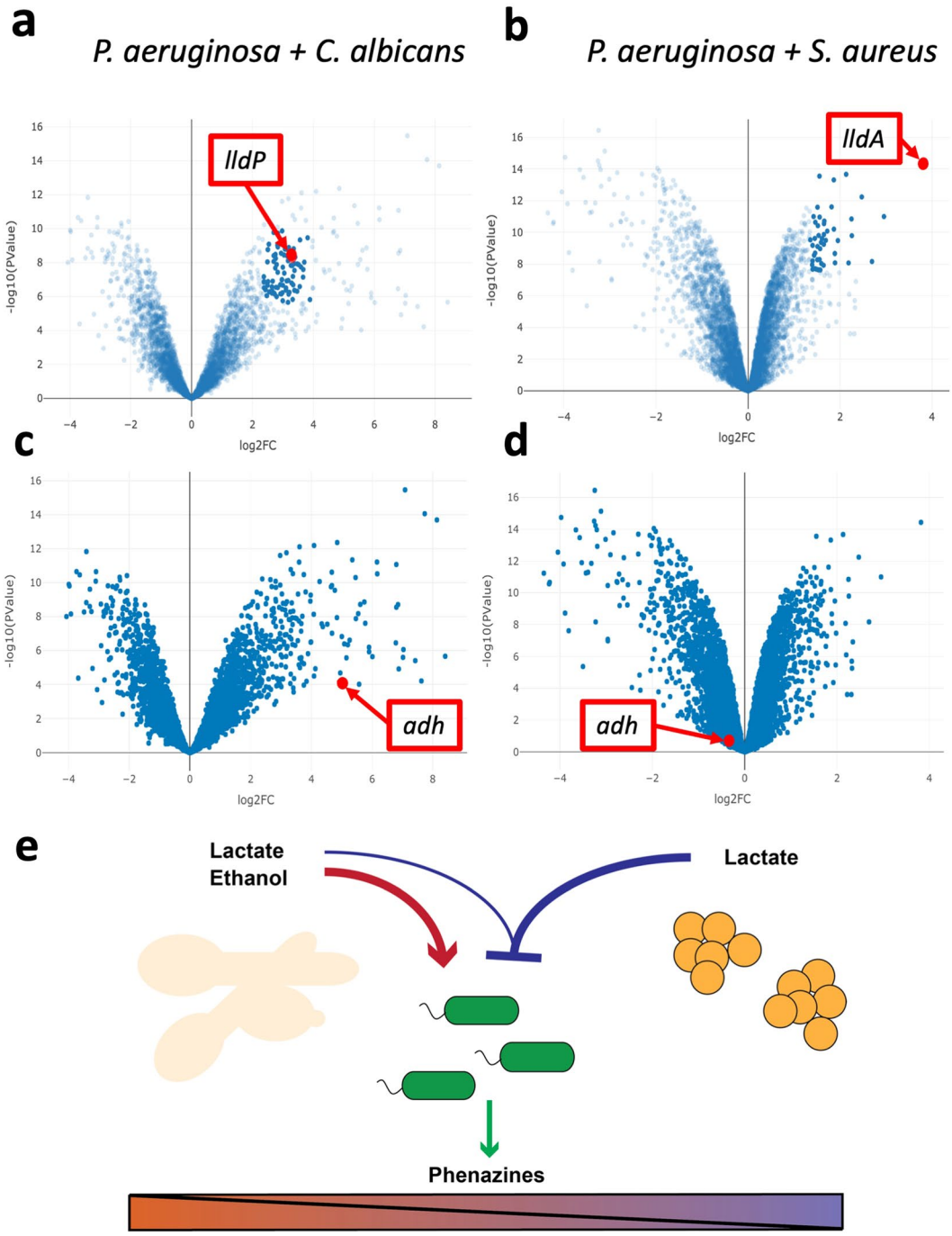

Fig. 5 (**a**) P. aeruginosa upregulates the expression of lactate permease lldP (red point) and other lactate metabolism genes including lactate dehydrogenases (present in the cluster of dark blue points near lldP) in co-culture with C. albicans. (**b**) Similarly, lactate dehydrogenase lldA (red point) and other lactate metabolism genes (included in dark blue points near lldA) are upregulated in co-culture with S. aureus as well. (**c**) P. aeruginosa upregulated alcohol dehydrogenase adh in co-culture with C. albicans (**d**) but not in co-culture with S. aureus. (**e**) In complex co-culture P. aeruginosa will have to integrate multiple signals such as the positive influence of ethanol and a possible negative influence of lactate that converge to influence phenazine production. After CF-Seq exploratory analysis, our hypothesis is that the presence of ethanol will supersede that of lactate to promote phenazine production.

repressive effect on phenazine production. In common laboratory mono-culture conditions, *P. aeruginosa* can produce an abundance of phenazines during exponential and stationary phase growth and thus conditions in which *P. aeruginosa* does not produce phenazines could contain repressive stimuli. Interestingly, given that lactate is oxidized by *P. aeruginosa*, catalyzed by lactate dehydrogenases including *lldA* and *lldD*[49,50], and phenazines

are highly redox active, tight control over phenazine production in the presence of lactate may be important for metabolic efficiency. This hypothesis could easily be tested in the lab by the addition of sub-lethal concentrations of ethanol to *P. aeruginosa* and *S. aureus* co-culture and measuring phenazine biosynthesis to test the hypothesis that phenazine production would increase.

Importantly, CF-Seq facilitated the re-analysis of public data that led to the development of a hypothesis in approximately 30 minutes. By contrast, the process of identifying these experiments, downloading the data, performing comparisons, and generating figures by hand would have taken approximately 16 hours, based on similar exploratory analyses that we have performed previously.

Case study #3: Examining superoxide dismutase response in *Staphylococcus aureus* under a variety of clinically relevant conditions. Dr. Liviu Cengher, Cheung Laboratory, Geisel School of Medicine.

Staphylococcus aureus is a human commensal and opportunistic pathogen that contributes to a wide range of diseases – from skin and soft-tissue disorders to respiratory diseases like cystic fibrosis[51]. Disease is mediated by several *S. aureus* virulence factors that are produced in response to environmental cues, and which play a wide range of roles[52]. Two-component systems (TCS) are important regulatory factors that have paired sensing and regulatory peptides that respond to environmental and host cues[53,54]. The *SaeR/S* TCS senses reactive oxygen species (ROS) and regulates responses that counteract and inhibit ROS production by the human immune system. For example, activation of the TCS may lead to enhanced expression of virulence factors superoxide dismutase *sodA* and *sodM*[55].

In this case study we investigated *sodA* and *sodM* expression across experiments with different bacterial strains and treatments to explore similarities and differences in ROS response. Specifically, we compared *sodA* and *sodM* expression in conditions likely to be present in the CF lung to identify conditions that upregulate one and/or both of the two genes. To start, we evaluated the effect of *S. aureus* co-culture with *P. aeruginosa* (vs. *S. aureus* in monoculture, GEO: GSE122048)[47]. Co-occurrence of *P. aeruginosa* and *S. aureus* is frequent in a hospital setting, and tends to induce a fermentative state in *S. aureus*[56,57]. Both *sodA* and *sodM* were upregulated in these conditions [Fig. 6a,b]. CF-Seq analysis of the transcriptome of 'persister cells' primed to survive (predominantly ROS mediated) killing after residing inside of immune system macrophages (GEO: GSE139659)[55,58] revealed that *sodM* was upregulated in the 'persister cells' that resisted killing by the immune system, though *sodA* was not [Fig. 6c,d]. Since *sodM* was upregulated in common between these two studies, it would be worth re-examining both conditions in tandem: subjecting *S. aureus* to bacterial co-culture with *P. aeruginosa* to see if this induces a persister-like phenotype in *S. aureus*. Given that both conditions – persistence within host cells and co-infection with *Pseudomonas* – may be present at once in an individual with CF, such experiments would paint a fuller picture of the *S. aureus* transcriptional state during an infection.

Furthermore, we also identified a study where treatment with apicidin, an antibiotic known to inhibit bacterial quorum sensing, led to downregulation of *sodA* and relatively low levels of *sodM* expression [Fig. 6e,f][59]. We might compare *sodA* downregulation in this study with the co-culture study (GEO: GSE122048). There would be interesting therapeutic implications for future experiments that determine the outcome of combining co-culture conditions (upregulating *sod* genes) with antibiotic-induced quorum sensing inhibition (downregulation/low expression of *sod* genes) to see which effect dominates. In addition, one might examine conditions which could favor *sodM* expression over *sodA* expression, like the availability of iron and manganese in co-culture and in polymicrobial infections[60,61]. Normally the analysis performed in this case study would necessitate a close reading of multiple published articles and require deciphering often unhelpful supplemental data tables. Finding relevant experiments and performing subsequent analysis would involve many hours of work. Using CF-Seq, useful results were found within approximately 10 minutes.

Discussion

As the user stories demonstrate, the CF-Seq application provides value to CF pathogen researchers in several ways. First, CF-Seq allows rapid analysis of numerous datasets, reducing the time of analysis from days in some cases to minutes. Second, multiple CF pathogens can be analyzed including bacteria such as *Pseudomonas aeruginosa* and *Staphylococcus aureus*, and fungi such as *Aspergillus fumigatus* and *Candida albicans*. Third, the R scripts underlying the application (publicly available in our Git Repository: https://github.com/samlo777/cf-seq.git) not only allow for rapid analysis of the CF pathogens currently included in the application, but may be repurposed to study other microbes relevant to other diseases. Furthermore, the annotation files provided in the Git repository (linking various gene identifiers and functional annotations like KEGG and GO terms) can be downloaded and utilized by CF researchers outside of the application, allowing them to perform functional analyses of their own datasets much more efficiently. Fourth, CF-Seq affords researchers a better understanding of prior CF pathogen experiments by revealing experimental parameters – details on strain, media, treatment, and gene perturbation – that have been tested in the past. With the ability to filter studies based on these parameters, users may identify the set of experiments that relate to their own specific interests and capabilities, filling knowledge gaps that they notice in the field of research. Not only does the application make prior studies more visible and accessible, but also makes their individual samples and the expression levels of individual genes possible to investigate more closely. While any given publication tends to emphasize the differential expression of just a few relevant genes to tell a concise and cohesive biological story, the CF-Seq application allows users to explore the expression of genes that may not have been of interest to the initial study authors but are of interest to the users themselves.

The ability to discern the whole field of prior experiments in minutes without slowly trawling through online databases like GEO is a tantalizing prospect. As it stands, the application serves as a valuable tool for validating existing hypotheses and generating new ones to test. That said, efforts are still ongoing to expand the application

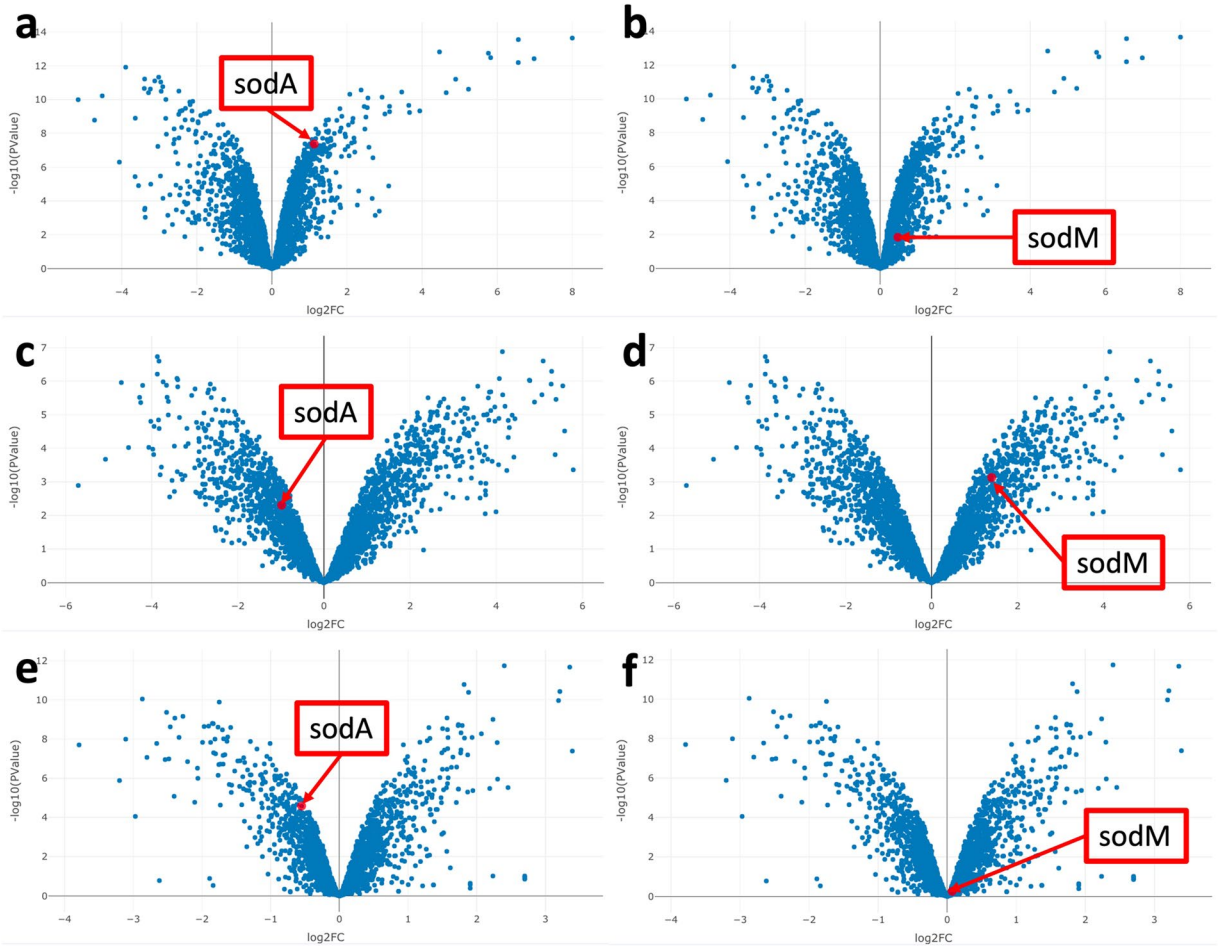

Fig. 6 Expression of virulence factors sodA and sodM in S. aureus tends to diverge under different experimental conditions. Volcano plots of all genes are shown to demonstrate the expression values of sodA and sodM relative to other genes detected. (**a,b**) In co-culture with P. aeruginosa, both sodA and SodM expression are upregulated, sodA to a much greater extent. (**c,d**). In 'persister cells', the expression pattern was quite different: sodM expression was more markedly upregulated while sodA expression was downregulated (**e,f**). Finally, exposure to apicidin was found to induce downregulation of sodA, but no significant change in sodM. In all cases, aside from sodM expression in (**f**) sodA and sodM were differentially expressed to a statistically significant degree ($p < 0.05$).

– adding older microarray studies to the compendium of data and making efforts to gather count table data for RNA-Seq studies in which count tables have not yet been provided directly by the authors as supplemental information in GEO. Additional RNA-Seq studies may be gathered by taking advantage of pipelines built to convert FASTQ sample files in GEO into count tables amenable to analysis by edgeR. For example, we may employ the pipeline recently developed by Doing *et al.* (2022) to create a compendium of *P. aeruginosa* datasets, modifying it such that its use extends to other CF pathogens of interest[62]. We may also take advantage of crowd-sourced metadata curation approaches like that of Wang *et al.* (2016), in which participants were recruited to help identify studies in GEO involving gene or drug perturbations, or comparison of normal and diseased tissue[63]. Crowdsourcing curation efforts would make the process of adding additional study data to the application more efficient and speed up the inclusion of new studies.

Though the application is not designed to compare studies head-to-head, the intrepid researcher could download the results of multiple studies and perform their own meta-analysis. That said, researchers need to be cautious of batch effects when comparing independent studies – as conditions ranging from the unique instruments of a given laboratory to the environmental conditions on the day of the experiment may impact the results. Meta-analysis should be conducted carefully with the application of appropriate statistical methods for batch effect correction[64,65]. Readers should also keep in mind that there is significant variation in the RNA-sequencing platforms (e.g., Illumina, PacBio, Oxford Nanopore) and the approach to pre-processing of data utilized by study authors across datasets. These factors further complicate comparison across studies. As an example of pre-processing differences, the study authors may choose to exclude non-coding RNA transcripts from the raw count data (though in studies where they are not excluded, the CF-Seq application allows users to assess their differential expression). Furthermore, the choice of reference strain (and genome) by the dataset

authors to map sequence reads to transcript IDs and produce raw count data introduces further differences between studies of the same species.

Most datasets currently featured in the application were generated by short-read cDNA sequencing on Illumina platforms. Over time, we will update the application with additional CF pathogen studies as they become available, including long-read sequencing datasets (typically generated by PacBio or Oxford Nanopore platforms) and even direct RNA-seq datasets. These more recent, improved RNA-Seq approaches eliminate some of the errors and biases in the data gathered by traditional short-read RNA-seq methods[66,67]. As the field of CF pathogen research evolves, and the methods it employs changes, the CF-Seq application will evolve as well.

In sum, the application sheds light on the value of automated bioinformatic analysis for researchers of all backgrounds. Performing differential expression analysis is by no means a feasible task for those lacking a computational background, and even for those who have such a background, analysis is still quite time-consuming (as the authors of the user stories note). Not only does the CF-Seq application save time and provide detailed statistical analysis, but it also serves a didactic purpose for those who have less experience working with transcriptomic data – demonstrating what differential expression analysis looks like and how it may be interpreted. Tools such as CF-Seq, and the other data re-analysis applications cited throughout this publication, demonstrate the immense value of bioinformatic tools for scientific research.

Providing CF pathogen researchers a more detailed view of the prior experiments conducted in their own domain will make research more coordinated, systematic, and efficient. The CF-Seq application allows users to see exactly what combinations of experimental factors have been assessed thus far, and take logical, incremental steps – investigate a new treatment, a new mutation, a new growth medium, or some combination thereof – to test novel experimental hypotheses and improve understanding of pathogen behavior. For the field of bioinformatics specifically, such an application helps demonstrate the value and enhance appreciation for both data re-analysis and the tools that enable it. More generally, applications like CF-Seq help democratize the research process, allowing all scientists, regardless of specialization, to set their minds at work determining where research should go next.

Methods

Data extraction. The CF-Seq application currently includes 128 RNA-Seq studies of 13 CF pathogens. All studies can be found in NCBI's Gene Expression Omnibus (GEO). Before incorporating studies into the application, the landscape of CF pathogen studies in GEO was surveyed. Clinically relevant pathogens of interest were chosen based on the cystic fibrosis literature (their relevance, supported by clinical and laboratory studies, is documented in Supplemental Table S1). The set of all RNA-sequencing studies for each of these pathogens was identified in GEO by querying the database of GEO datasets by pathogen name (e.g., *Pseudomonas aeruginosa*, *Staphylococcus aureus*, etc.), filtering studies to include only those that constituted "expression profiling by high throughput sequencing" (in GEO, this corresponds to 'RNA sequencing'), and selecting the pathogen of interest specifically in the 'organism' field. This final step excludes datasets that constitute transcriptomic profiles of human cells, or cells of some other organism, exposed to the pathogen of interest.

For practical reasons, only studies with certain attributes are included in this release of the compendium. The application is limited to studies where: A) a count table was provided in the supplemental files associated with the study in GEO, B) that count table was in a tabular format (.csv, .xlsx, .txt) so that it could be loaded into R with the read.table() or read.csv() functions, C) sample groups were clearly distinguishable such that it was possible to perform differential expression analysis, and D) the count table included raw counts and not normalized counts (edgeR and other differential expression analysis packages require raw counts to perform analysis). Efforts to circumvent some of these limitations and add more studies into the application are discussed in the Discussion section of this manuscript.

Data cleaning and storage. For studies that met the criteria for inclusion in the application, each count table was subjected to the following formatting protocol. Count tables downloaded directly from GEO were re-structured, if necessary, so that the first column of the table included gene names, and all subsequent columns contained raw read data for each experimental sample. In addition to count tables, two other data files were constructed for each study. The first file is a design matrix which delineates experimental samples by condition (e.g., control, treatment group X, treatment group Y) and lists the number of replicates for each condition. This design matrix is a requirement for differential expression analysis with edgeR. The second file, labeled 'additional metadata', includes manually gathered metadata on the strain(s), media, treatment conditions, and genes perturbed in each study, whenever applicable. Collecting this data enables filtering of studies by experimental conditions within the application.

To build the application, all data files – count tables, design matrices, and additional metadata – were deposited in a local directory of folders, with a single folder for each species, and sub-folders within each species for the three types of data files (count table, design matrix, additional metadata). A copy of this directory structure can be found in the Git Repository associated with this publication [https://github.com/samlo777/cf-seq.git], so that any reader may download the data and/or use it to run the Shiny application on their own computer if they so choose.

Code development approach. The CF-Seq application code was developed in discrete modules to make testing as straight-forward as possible. Each of the application's interactive features (filtering studies, selecting a study, choosing experimental comparisons to analyze, etc.) were developed in a hierarchical fashion: the code was first tested to ensure that it worked properly for a single study, then adjusted and generalized such that it worked for a single species, and ultimately for all species included in the application.

The application code is broken down into 3 files. The first, named 'app.R', contains the functional code for the application. This file houses the UI code (which dictates the appearance of the application), and the server code (which provides functional code for all the drop-down menus, tables, and output figures) as two separate blocks. Both the name of this file, and the two-section structure, are an essential requirement of all Shiny applications. In addition, another code file, labeled 'Data Setup.R', was generated to load in all the study data and compress it into an easily accessible data structure (a list of lists) accessible to the code in the app.R file. In addition to loading in the count tables, design matrices, and additional metadata, this code file also contains blocks of code that perform differential expression analysis – and deposit the outputs of this analysis (including tables of fold changes, p values, and counts per million for each gene) into the list of lists object alongside their respective studies. The third code file, labeled 'Annotation Data.R', contains code that programmatically accesses data from the KEGG[68–70], UniProt[71], and COG[72] databases and structures that data such that the genes of each species are linked to their respective KEGG biological pathway identifiers, GO Terms, and COG categories (if available).

The application takes advantage of several publicly available, open-source R packages. Alongside the 'shiny' package[73] (which is essential for all R Shiny applications), the 'shinydashboard'[74] package was used to provide a UI template, with several tabs for different application components. 'shinyjs'[75] was used to develop some of the more complicated application features (e.g., data tables with interactive buttons) that require JavaScript code to run. The 'DT'[76] package was employed to create searchable and filterable tables. The package 'plotly'[77] was used to generate interactive volcano plots and MA plots to represent differential expression analysis results, and the differential gene expression analysis itself was performed with the 'edgeR'[28,29] package. The KEGGREST[78] and UniProt.ws[79] packages were used to download functional annotations (KEGG pathways, GO terms, COG categories) to pair with certain amenable studies in the application. Finally, the 'tidyverse'[80] suite of packages, including 'stringr'[81] for string manipulation, were used throughout the application code to manipulate data structures.

Validation: beta testing protocol. To ensure that the study data and metadata loaded into the application recapitulated the data present in GEO, and that all application features worked as expected, a beta testing protocol was established. Three of the paper co-authors, each possessing either domain knowledge in CF microbiology or bioinformatics, were recruited to test different segments of the application: (1) the ability to filter studies based on experimental characteristics, (2) the ability to view detailed metadata for each individual study, and (3) the ability to perform and visualize differential expression analysis. The beta testing protocol was guided by a series of requirements tables that listed out all the features to be validated (beta testers were instructed to indicate Y/N if a feature worked as expected and provide notes if it did not). These tables are included for reference in the supplemental material [Supplemental Tables S5–S7]

After all components of the application were tested, any features that did not work properly were fixed – and additional improvements were made to enhance the usability of the application based on beta tester feedback. Furthermore, after the bugs identified in beta testing were fixed, a second round of review was undertaken to ensure that study metadata accurately reflected the true study metadata in GEO. One at a time, each study in the application was referenced back to GEO to ensure that none of the manually curated metadata was missing or incorrect.

Project documentation. Documentation for the CF-Seq application can be found in several locations. Users are presented with a user manual when they first open the application. Further guidance on using the application can also be found in the form of the README file in the Git repository associated with this project [https://github.com/samlo777/cf-seq.git].

Data availability

All data – including count tables derived from GEO, and manufactured design matrices and metadata tables – are available in the Git repository [https://github.com/samlo777/cf-seq.git].

Code availability

All CF-Seq code is open source and has been made available for use on GitHub under the MIT License [https://github.com/samlo777/cf-seq.git].

The application is also hosted on a server maintained by Dartmouth College and is accessible at the following web link [http://scangeo.dartmouth.edu/CFSeq/].

In its current version, CF-Seq utilizes the following R package versions: *shiny* (1.6.0), *shinydashboard* (0.7.1), *shinyjs* (2.0.0), *DT* (0.19.1), *plotly* (4.9.4.1), *ggplot2* (3.3.5), *edgeR* (3.34.1), *KEGGREST* (1.32.0), *UniProt.ws* (2.32.0), *tidyverse* (1.3.1), *stringr* (1.4.0).

Received: 9 March 2022; Accepted: 25 May 2022;
Published online: 16 June 2022

References

1. About cystic fibrosis. *Cystic Fibrosis Foundation* https://www.cff.org/intro-cf/about-cystic-fibrosis (n.d.).
2. Kälin, N., Claaß, A., Sommer, M., Puchelle, E. & Tümmler, B. ΔF508 CFTR protein expression in tissues from patients with cystic fibrosis. *J Clin Invest* **103**, 1379–1389 (1999).
3. Painter, R. G. *et al*. CFTR expression in human neutrophils and the phagolysosomal chlorination defect in cystic fibrosis. *Biochemistry* **45**, 10260–10269 (2006).
4. Tousson, A., Van Tine, B. A., Naren, A. P., Shaw, G. M. & Schwiebert, L. M. Characterization of CFTR expression and chloride channel activity in human endothelia. *American Journal of Physiology-Cell Physiology* **275**, C1555–C1564 (1998).
5. Kelly, T. & Buxbaum, J. Gastrointestinal manifestations of cystic fibrosis. *Dig Dis Sci* **60**, 1903–1913 (2015).
6. Lyczak, J. B., Cannon, C. L. & Pier, G. B. Lung infections associated with cystic fibrosis. *Clinical Microbiology Reviews* (2002).
7. Bonfield, T. L. *et al*. Inflammatory cytokines in cystic fibrosis lungs. *Am J Respir Crit Care Med* **152**, 2111–2118 (1995).

8. Kreda, S. M., Davis, C. W. & Rose, M. C. CFTR, Mucins, and mucus obstruction in cystic fibrosis. *Cold Spring Harb Perspect Med* **2**, a009589 (2012).
9. Hey, J. *et al*. Epigenetic reprogramming of airway macrophages promotes polarization and inflammation in muco-obstructive lung disease. *Nat Commun* **12**, 6520 (2021).
10. Hayden, H. S. *et al*. Fecal dysbiosis in infants with cystic fibrosis is associated with early linear growth failure. *Nature Medicine* **26**, 215–222 (2020).
11. Garg, M. & Ooi, C. Y. The enigmatic gut in cystic fibrosis: linking inflammation, dysbiosis, and the increased risk of malignancy. *Curr Gastroenterol Rep* **19**, 6 (2017).
12. Understanding changes in life expectancy. *Cystic Fibrosis Foundation* https://www.cff.org/managing-cf/understanding-changes-life-expectancy (n.d.)
13. Balfour-Lynn, I. M. & King, J. A. CFTR modulator therapies – effect on life expectancy in people with cystic fibrosis. *Paediatric Respiratory Reviews* (2020).
14. Dave, K. *et al*. Entering the era of highly effective modulator therapies. *Pediatric Pulmonology* **56**, S79–S89 (2021).
15. Ricotta, E. E., Prevots, D. R. & Olivier, K. N. CFTR modulator use and risk of nontuberculous mycobacteria positivity in cystic fibrosis, 2011–2018. *ERJ Open Res* **8**, 00724–02021 (2022).
16. Heltshe, S. L. *et al*. Pseudomonas aeruginosa in Cystic Fibrosis Patients with G551D-CFTR Treated With Ivacaftor. *Clin Infect Dis* **60**, 703–712 (2015).
17. Frost, F. J., Nazareth, D. S., Charman, S. C., Winstanley, C. & Walshaw, M. J. Ivacaftor Is associated with reduced lung infection by key cystic fibrosis pathogens. A cohort study using national registry data. *Ann Am Thorac Soc* **16**, 1375–1382 (2019).
18. Durfey, S. L. *et al*. Combining ivacaftor and intensive antibiotics achieves limited clearance of cystic fibrosis infections. *mBio* **12**, e03148–21 (2021).
19. Veziris, N. *et al*. Non-tuberculous mycobacterial pulmonary diseases in France: an 8 years nationwide study. *BMC Infect Dis* **21**, 1165 (2021).
20. Johansen, M. D., Herrmann, J.-L. & Kremer, L. Non-tuberculous mycobacteria and the rise of *Mycobacterium abscessus*. *Nat Rev Microbiol* **18**, 392–407 (2020).
21. Silveira, C. B. *et al*. Multi-omics study of keystone species in a cystic fibrosis microbiome. *Int J Mol Sci* **22**, 12050 (2021).
22. Pust, M.-M. & Tümmler, B. Bacterial low-abundant taxa are key determinants of a healthy airway metagenome in the early years of human life. *Computational and Structural Biotechnology Journal* **20**, 175–186 (2022).
23. O'Toole, G. A. *et al*. Model systems to study the chronic, polymicrobial infections in cystic fibrosis: current approaches and exploring future directions. *mBio* (2021).
24. Widder, S. *et al*. Association of bacterial community types, functional microbial processes and lung disease in cystic fibrosis airways. *ISME J* 1–10 (2021).
25. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417–419 (2017).
26. R Core Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, 2018).
27. RStudio Team. *RStudio: Integrated Development for R*. (RStudio, Inc., 2019).
28. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
29. Chen, Y., Lun, A. T. L. & Smyth, G. K. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Res* **5**, 1438 (2016).
30. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).
31. Wilkinson, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
32. Kumar, P. *et al*. MetaRNA-Seq: an interactive tool to browse and annotate metadata from RNA-Seq studies. *Biomed Res Int* **2015**, 318064 (2015).
33. Alameer, A. & Chicco, D. GeoCancerPrognosticDatasetsRetriever: a bioinformatics tool to easily identify cancer prognostic datasets on Gene Expression Omnibus (GEO). *Bioinformatics* btab852 (2021).
34. Koeppen, K., Stanton, B. A. & Hampton, T. H. ScanGEO: parallel mining of high-throughput gene expression data. *Bioinformatics* **33**, 3500–3501 (2017).
35. Li, Z. *et al*. GAUGE-annotated microbial transcriptomic data facilitate parallel mining and high-throughput reanalysis to form data-driven hypotheses. *mSystems* **6**, e01305–20 (2021).
36. Latgé, J.-P. & Chamilos, G. *Aspergillus fumigatus* and aspergillosis in 2019. *Clinical Microbiology Reviews* (2019).
37. Ross, B. S., Lofgren, L. A., Ashare, A., Stajich, J. E. & Cramer, R. A. *Aspergillus fumigatus* in-host HOG pathway mutation for cystic fibrosis lung microenvironment persistence. *mBio* (2021).
38. Keown, K., Reid, A., Moore, J. E., Taggart, C. C. & Downey, D. G. Coinfection with *Pseudomonas aeruginosa* and *Aspergillus fumigatus* in cystic fibrosis. *European Respiratory Review* **29**, (2020).
39. *2019 Patient Registry Annual Data Report*. https://www.cff.org/sites/default/files/2021-10/2019-Patient-Registry-Annual-Data-Report.pdf (Cystic Fibrosis Foundation, 2019).
40. Camus, L., Briaud, P., Vandenesch, F. & Moreau, K. How bacterial adaptation to cystic fibrosis environment shapes interactions between *Pseudomonas aeruginosa* and *Staphylococcus aureus*. *Frontiers in Microbiology* **12** (2021).
41. Delhaes, L. *et al*. The Airway microbiota in cystic fibrosis: a complex fungal and bacterial community—implications for therapeutic management. *PLOS ONE* **7**, e36313 (2012).
42. Jean-Pierre, F., Vyas, A., Hampton, T. H., Henson, M. A. & O'Toole, G. A. One versus many: polymicrobial communities and the cystic fibrosis airway. *mBio* **12**, e00006–21 (2021).
43. Khanolkar, R. A. *et al*. Ecological succession of polymicrobial communities in the cystic fibrosis airways. *mSystems* (2020).
44. Quinn, R. A. *et al*. Ecological networking of cystic fibrosis lung infections. *NPJ Biofilms Microbiomes* **2**, 4 (2016).
45. Doing, G., Koeppen, K., Occipinti, P., Harty, C. E. & Hogan, D. A. Conditional antagonism in co-cultures of *Pseudomonas aeruginosa* and *Candida albicans*: an intersection of ethanol and phosphate signaling distilled from dual-seq transcriptomics. *PLoS Genet* **16**, e1008783 (2020).
46. Bisht, K., Baishya, J. & Wakeman, C. A. *Pseudomonas aeruginosa* polymicrobial interactions during lung infection. *Current Opinion in Microbiology* **53**, 1–8 (2020).
47. Tognon, M., Köhler, T., Luscher, A. & van Delden, C. Transcriptional profiling of *Pseudomonas aeruginosa* and *Staphylococcus aureus* during *in vitro* co-culture. *BMC Genomics* **20**, 30 (2019).
48. Chen, A. I. *et al*. *Candida albicans* ethanol stimulates *Pseudomonas aeruginosa* WspR-controlled biofilm formation as part of a cyclic relationship involving phenazines. *PLoS Pathog* **10**, e1004480 (2014).
49. Lin, Y.-C., Cornell, W. C., Jo, J., Price-Whelan, A. & Dietrich, L. E. P. The Pseudomonas aeruginosa complement of lactate dehydrogenases enables use of d- and l-lactate and metabolic cross-feeding. *mBio* **9**, e00961–18 (2018).
50. Jo, J., Price-Whelan, A., Cornell, W. C. & Dietrich, L. E. P. Interdependency of respiratory metabolism and phenazine-associated physiology in *Pseudomonas aeruginosa* PA14. *Journal of Bacteriology* **202**, e00700–19 (2020).
51. Lowy, F. D. *Staphylococcus aureus* infections. *New England Journal of Medicine* **339**, 520–532 (1998).

52. Cheung, A. L., Bayer, A. S., Zhang, G., Gresham, H. & Xiong, Y.-Q. Regulation of virulence determinants *in vitro* and *in vivo* in *Staphylococcus aureus*. *FEMS Immunology & Medical Microbiology* **40**, 1–9 (2004).
53. Haag, A. F. & Bagnoli, F. The role of two-component signal transduction systems in *Staphylococcus aureus* virulence regulation. in *Staphylococcus aureus: Microbiology, Pathology, Immunology, Therapy and Prophylaxis* (eds. Bagnoli, F., Rappuoli, R. & Grandi, G.) 145–198 (Springer International Publishing, 2017).
54. Wu, S., Lin, K., Liu, Y., Zhang, H. & Lei, L. Two-component signaling pathways modulate drug resistance of *Staphylococcus aureus* (Review). *Biomed Rep* **13**, 5 (2020).
55. Guerra, F. E. *et al*. *Staphylococcus aureus* SaeR/S-regulated factors reduce human neutrophil reactive oxygen species production. *J Leukoc Biol* **100**, 1005–1010 (2016).
56. Filkins, L. M. *et al*. Coculture of *Staphylococcus aureus* with *Pseudomonas aeruginosa* drives *S. aureus* towards fermentative metabolism and reduced viability in a cystic fibrosis model. *Journal of Bacteriology* (2015).
57. Fischer, A. J. *et al*. Sustained coinfections with *Staphylococcus aureus* and *Pseudomonas aeruginosa* in cystic fibrosis. *Am J Respir Crit Care Med* **203**, 328–338 (2021).
58. Peyrusson, F. *et al*. Intracellular *Staphylococcus aureus* persisters upon antibiotic exposure. *Nat Commun* **11**, 2200 (2020).
59. Parlet, C. P. *et al*. Apicidin attenuates MRSA virulence through quorum-sensing inhibition and enhanced host defense. *Cell Rep* **27**, 187–198.e6 (2019).
60. Garcia, Y. M. *et al*. A superoxide dismutase capable of functioning with iron or manganese promotes the resistance of *Staphylococcus aureus* to calprotectin and nutritional immunity. *PLOS Pathogens* **13**, e1006125 (2017).
61. Lalaouna, D. *et al*. RsaC sRNA modulates the oxidative stress response of *Staphylococcus aureus* during manganese starvation. *Nucleic Acids Research* **47**, 9871–9887 (2019).
62. Doing, G. *et al*. Computationally efficient assembly of a *Pseudomonas aeruginosa* gene expression compendium. Preprint at https://www.biorxiv.org/content/10.1101/2022.01.24.477642v1 (2022).
63. Wang, Z. *et al*. Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nat Commun* **7**, 12846 (2016).
64. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics* **2**, lqaa078 (2020).
65. Tran, H. T. N. *et al*. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology* **21**, 12 (2020).
66. Amarasinghe, S. L. *et al*. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology* **21**, 30 (2020).
67. Depledge, D. P. *et al*. Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nat Commun* **10**, 754 (2019).
68. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28**, 27–30 (2000).
69. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci* **28**, 1947–1951 (2019).
70. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* **49**, D545–D551 (2020).
71. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**, D480–D489 (2021).
72. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res* **43**, D261–D269 (2015).
73. Chang, W. *et al*., shiny: Web Application Framework for R. *CRAN* https://cran.r-project.org/web/packages/shiny/index.html (2021).
74. Chang, W. & Borges Ribeiro, B. shinydashboard: Create Dashboards with 'Shiny'. *CRAN* https://cran.r-project.org/web/packages/shinydashboard/index.html (2018).
75. Attali, D. shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds. *CRAN* https://cran.r-project.org/web/packages/shinyjs/index.html (2021).
76. Xie, Y. *et al*. DT: A Wrapper of the JavaScript Library 'DataTables'. *CRAN* https://cran.r-project.org/web/packages/DT/index.html (2022).
77. Sievert, C. *Interactive Web-Based Data Visualization with R, plotly, and shiny*. (Chapman and Hall/CRC, 2020).
78. Tenenbaum D, Maintainer B. KEGGREST: Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG). R package version 1.36.0. *Bioconductor* https://bioconductor.org/packages/release/bioc/html/KEGGREST.html (2022).
79. Carlson M, Maintainer B. UniProt.ws: R Interface to UniProt Web Services. R package version 2.36.0. *Bioconductor* https://bioconductor.org/packages/release/bioc/html/UniProt.ws.html (2022).
80. Wickham, H. *et al*. Welcome to the Tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).
81. Wickham, H. stringr: Simple, Consistent Wrappers for Common String Operations. *CRAN* https://cran.r-project.org/web/packages/stringr/index.html (2019).

Acknowledgements

Support for these studies was provided by the Cystic Fibrosis Foundation (grants STANTO19R0 – CFF and CRAMER19GO – CFF), the NIH (grants P30 DK117469, R01 HL151385, and R01AI146121), and The Flatley Foundation.

Author contributions

S.N. wrote the publication except for the user stories, and conceived of, developed, and tested the CF-Seq application. T.H. provided valuable guidance and inspiration through the entire application development and publication writing process. In addition, T.H. helped gather gene annotations to allow more sophisticated analysis of studies in the application. K.K. provided helpful guidance on app development and handled hosting of the app on the online server. L.C., C.P. and A.L. all helped with beta testing the application. L.C., C.P. and G.D. provided user stories demonstrating how the application could be used to test wet bench hypotheses. B.S. provided domain expertise (specifically, on *P. aeruginosa*), guidance on publication, and primary funding for the project. A.C., D.H. and R.C. also provided domain expertise and guidance. All authors reviewed drafts of the publication and provided feedback.

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/s41597-022-01431-1.

Correspondence and requests for materials should be addressed to B.A.S.