There is no hiding if you Seq: recent breakthroughs in Pseudomonas aeruginosa research revealed by genomic and transcriptomic next-generation sequencing

Valli, Richard Xavier Etienne; Lyng, Mark; Kirkpatrick, Clare

Published in:
Journal of Medical Microbiology

DOI:
10.1099/jmm.0.001135

Publication date:
2020

Document version
Accepted manuscript

Document license
CC BY

Citation for published version (APA):

Terms of use
This work is brought to you by the University of Southern Denmark through the SDU Research Portal. Unless otherwise specified it has been shared according to the terms for self-archiving. If no other license is stated, these terms apply:

• You may download this work for personal use only.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying this open access version

If you believe that this document breaches copyright please contact us providing details and we will investigate your claim. Please direct all enquiries to puresupport@bib.sdu.dk

Download date: 11. Jan. 2021
Abstract
The advent of next-generation sequencing technology has revolutionized the field of prokaryotic genetics and genomics by allowing interrogation of entire genomes, transcriptomes and global transcription factor binding profiles. As more studies employing these techniques have been performed, the state of the art regarding prokaryotic gene regulation has developed from the level of individual genes to genetic regulatory networks and systems biology. When applied to bacterial pathogens, particularly valuable insights have been gained into the regulation of virulence-associated genes, their relative importance to bacterial survival in planktonic, biofilm or host infection scenarios, antimicrobial resistance and the molecular details of host-pathogen interaction. This Review outlines some of the latest developments and applications of next-generation sequencing techniques to the genus Pseudomonas, with particular focus on opportunistic human pathogen Pseudomonas species, the biological insights gained from them and the future directions in which this field could develop.

Keywords
Pseudomonas; Next-generation sequencing; RNA-Seq; Transposon insertion sequencing; ChIP-Seq; host-pathogen interaction

Abbreviations
ChIP-Seq, Chromatin Immunoprecipitation sequencing; NGS, next-generation sequencing; RNA-BP, RNA-binding protein; sRNA, small non-coding RNA; TIS, Transposon Insertion Sequencing
Introduction

The development of microarray technology and its application to global gene expression analysis in bacteria in the late 1990s (1, 2) marked a paradigm shift from the study of single-gene expression to interrogation of an entire transcriptome in one experiment. However, such experiments still required many a priori assumptions (primarily from selection of probes to include in the array) and suffered from technical hindrances such as a limited dynamic range of detection of expression levels and non-specific or cross-hybridization between samples. In addition to these technical issues, it was not usually possible to directly compare the results of one microarray to another unless complex normalization or statistical transformation methods were used, and experiments were limited to two-dimensional comparisons of control to test condition. The evolution of high-throughput sequencing, especially the later iterations known as next-generation sequencing (NGS) which lacked the primer-dependence of the classical Sanger sequencing method, revolutionized the field by allowing direct quantification of expression levels based on digital counts of NGS “reads” (3). With this read-out, the detection range of transcriptomic experiments was no longer limited by the biophysical constraints of probe fluorescence ratios but only by the total number of sequencing reads obtained (“sequencing depth”) allowing a far greater dynamic range of accurate quantification and a reduced signal-to-noise ratio. Moreover, since all RNA transcripts were represented in the sequenced samples, probe selection bias was no longer a factor, enabling the discovery of many uncharacterized or unannotated RNA molecules, such as non-coding RNAs or mRNAs of previously unknown small proteins.

Widespread adoption of this technology, given the popular name RNA-Seq (4), led to an exponential increase in the number of whole-transcriptome experiments reported in the literature in both prokaryotic and eukaryotic model organisms from 2009 to the present day. In parallel, NGS-based applications were developed for other purposes than whole-transcriptome sequencing. Taking advantage of the independence of NGS from probes or sequencing primers, this technology was applied to DNA obtained from chromatin immunoprecipitation (ChIP-Seq) from specific transcription factors to gain a genome-wide overview of their direct regulons in eukaryotes (reviewed by Park (5)) and prokaryotes (reviewed by Myers et al. (6)), and to prokaryote-specific applications such as genome sequencing on microbial communities rather than pure cultures (7). More recently, NGS has been combined with transposon mutagenesis in four independent but functionally similar methods (reviewed by van Opijnen and Camilli (8)) to illuminate global conditional essentiality genotypes associated with bacterial survival under specific conditions, in particular virulence of bacterial pathogens.
Due to the ubiquitous nature of *Pseudomonas* species as environmental and human associated bacteria, and the interest in their ability to form biofilms (9, 10), they were adopted relatively early as model organisms for NGS-based experiments and continue to be used as such. In this Review, we focus on genomic and transcriptomic applications of NGS that have led to recent breakthroughs in *Pseudomonas* research, especially those which have uncovered novel aspects of *Pseudomonas* regulatory networks and virulence-associated genotypes and phenotypes.

The dawn of the unbiased transcriptomics era

By applying RNA-Seq, it was possible for the first time to interrogate the full transcriptome without reliance on annotation-dependent probe design and the biases associated with it. As such, it was quickly adopted by the *Pseudomonas* research community to define the transcriptional responses of *P. aeruginosa* to conditions which can be encountered during infection, for example biofilm vs planktonic growth (11), growth at 28°C vs 37°C (12) or presence of human respiratory mucus (13). While these studies gave important insights at the time, such as the existence of previously unknown small non-coding RNAs that were differentially induced under the conditions tested, these binary comparisons of a single test condition to an untreated control could not replicate the multiple environmental differences experienced by bacteria grown in vitro and bacteria growing in the context of human infection, and accordingly very little overlap was observed between the differentially expressed genes found in the different studies. In order to obtain a more detailed view of the relative importance of genetic and environmental contributions to the global transcriptomic profile, a large-scale study of 151 *P. aeruginosa* clinical isolates grown in identical conditions alongside the laboratory strain PA14 grown under 14 different conditions was performed (14). Comparison by principal component analysis of the RNA-Seq data sets showed that the transcriptional profiles (and therefore phenotypic variations) depended more strongly on environmental stimuli than on genetic variation; the genetically diverse clinical isolates clustered together while the genetically identical laboratory strain transcriptomes diverged from each other to form environment-specific clusters. Moreover, when genetic variation was a determining factor, it was restricted to mutations within global transcriptional regulators, and the phenotypic variation resulted from the pleiotropic downstream effects of the mutated regulator on its target genes. It therefore became clear that in order to fully understand the mechanisms leading to expression of environment-specific transcriptional profiles, it would be necessary to define the regulons of the key transcription factors controlling virulence-related behaviors such as biofilm formation, motility, quorum sensing or expression of toxin secretion systems. This approach would characterize the network relationships between the different regulators and distinguish between direct and indirect regulation. RNA-Seq was used together with ChIP-Seq (see later) to identify direct and indirect
targets of virulence regulatory transcription factors such as AmpR (15) and AmrZ (16), which began to
deﬁne the full range of cellular processes that these factors provided regulatory input into (often more
than initially anticipated), and laid the groundwork for later studies that proﬁled the targets of multiple
transcription factors simultaneously (17, 18). Below, we highlight recent RNA-Seq studies using
Pseudomonas as model system which are of particular interest for the methodological innovations and/or
biological insights they provide. We discuss signiﬁcant advances in the Pseudomonas research ﬁeld that
have been obtained from ChIP-Seq and transposon insertion sequencing, alone or in combination with
RNA-Seq, and the future research directions in which these techniques may play a leading role.

Transcriptional changes during host-pathogen interaction

As RNA sample preparation methods advanced, it became possible to study transcriptomic proﬁles of
clinical isolates taken directly from human infection sites, allowing a better characterization of the
transcriptional response to growth in vivo during human infection. Consistent with the work of Dötsch et al.
(14) which had indicated that the environment of the bacteria is more inﬂuential over the transcriptional
proﬁle than their genotype, it was seen that that multiple genetically divergent P. aeruginosa lung isolates
displayed highly similar and genotype-independent transcriptional proﬁles (19, 20). A similar trend was
observed for P. aeruginosa clinical isolates taken from cystic ﬁbrosis patients’ lung expectorate, acute
wounds and chronic wounds compared to cultures grown under a large range of in vitro conditions. The in
vivo isolates had much more similarities in transcriptional proﬁles to each other than they did to any of the
in vitro cultures, even when they came from different wound sites, patients, geographical locations or
strain backgrounds (21). However, the gene expression pattern observed in pathogen during infection only
tells half of the story; an inescapable component of the environment of the pathogen comes from the
surrounding host cells and their response to the bacterial invasion.

During infection, both pathogen and host must adapt, and this translates into changes in gene expression.
Those changes are studied by “dual transcriptomics” or dual RNA-Seq, which consists of proﬁling RNA
expression simultaneously in the pathogen and the host (reviewed by Westermann et al. (22)). (28)(29)A
pair of recent studies have focused on P. aeruginosa infection in a mouse (23) and a zebrafish embryo (24)
model and provided novel insights into the infection process. In both cases, as expected, competition for
iron between host and pathogen was evident from the upregulation of iron scavenging genes in the
bacteria, as well as other virulence factors such as type 3 and type 6 secretion systems, but the responses
of the host organisms differed noticeably. The mouse response appeared to be much more narrowly
focused with relatively few changes in gene expression (hundreds, rather than thousands as observed in
the fish) and primarily directed toward iron sequestration, as changes were mostly in the transcripts of iron
binding factors, as well as other innate immune components. By contrast, the fish model had approximately one quarter of all its (protein coding) genes differentially regulated in the infection condition relative to control, and the major focus of the response appeared to be the activation of the innate immune system simultaneously with an intensive downregulation of protein translation, as ribosomal proteins, translation accessory factors and the 5.8S rRNA were among the most strongly downregulated transcripts. Although these studies differed in the infection route (acute lung infection in mouse, direct injection in zebrafish) and in the analytical methods used to process the RNA-Seq data, they have provided valuable information on how widely the adaptation process can differ between two different vertebrate hosts exposed to the same bacterium. As the zebrafish embryo model system in this study was used at a developmental stage when the innate immune system is functional but the adaptive immune response has not yet developed, while the mouse model employed adult mice with fully functioning adaptive and innate immune systems, comparison of the two models allows inferences to be made about the relative contributions of the two aspects of the vertebrate immune system to fighting *Pseudomonas* infection.

*Bacterial post-transcriptional regulation by small non-coding RNA (sRNA) and RNA-binding proteins (RNA-BPs)*

While the transcriptome provides a global profile of which genes are expressed, it cannot be assumed that the mRNA levels correlate perfectly with the protein levels in the cell. This is due to post-transcriptional regulation that can up- or down-regulate translation of the mRNA into protein, often mediated by sRNA binding to a partially or completely complementary mRNA molecule. The annotation-independence of RNA-Seq transcriptomics allowed these sRNAs to be detected without probing specifically for them, allowing a global view of condition-specific sRNA expression and defining their functions and targets. The most common (but not exclusive) mode of action of a sRNA is to base pair with its target mRNA molecule, often mediated by the conserved RNA chaperone Hfq, resulting in RNase recruitment and degradation of the duplex RNA. This negative regulation is often observed as a stress response to disadvantageous environmental conditions (reviewed by Hoe *et al.* (25)), as it allows for a faster response than regulation at the level of transcription initiation. However, some examples of sRNA binding leading to positive regulation have also been observed (reviewed by Papenfort and Vanderpool, (26)) and alternative modes of action such as decoy, sponge or titrating sRNAs have been recently discovered (reviewed by Kavita *et al.* (27)).

Several methods have been developed to study sRNA and RNA-BP interactions in bacteria (dRNA-Seq, pRNA-Seq, MAP-Seq, RIP-Seq, CLIP-Seq, CLASH, RIL-Seq) each one presenting different advantages and limitations (Figure 1, and reviewed extensively in (28)). dRNA-Seq and pRNA-Seq specifically allow the detection of transcription start sites and RNA cleavage sites, respectively, and have been used to
comprehensively map *P. aeruginosa* promoter and transcription start site locations, including for the entire suite of *P. aeruginosa* sRNAs (29). RIP-Seq and CLIP-Seq are used to study interactants of a specific RNA-BP by expressing an affinity tagged variant of the RNA-BP allowing it to be purified along with the RNAs bound to it. MAP-Seq allows the study of sRNA-mRNA binding by using a chimeric sRNA tagged with the MS2 RNA affinity tag, which enables purification of complexes via the MS2-binding protein. Finally, CLASH and RIL-Seq contain aspects of both since they use affinity enrichment of RNA-BP/RNA complexes to study RNA-RNA interactions and binding sites.

Due to the extensive conservation of the Hfq protein in bacteria and the observation that in *Pseudomonas* (30) (and other pathogens (25)) *hfq*-mutants are defective in virulence, a great deal of interest was directed towards identifying which sRNAs required Hfq for their activity. As a result, RNA-Seq methods such as those described above were developed to enrich for Hfq-bound sRNAs (and also applied to other RNA-BPs). However, these also led to the unexpected realization that Hfq could also interact with mRNAs, independently of chaperoning their interaction with a cognate sRNA. Two recent studies on RNA-BPs of *P. aeruginosa* revealed that the RNA-BPs Hfq and Crc (the carbon catabolite repression regulator) act co-transcriptionally with hundreds of nascent mRNA transcripts, many more than previously suspected (31, 32). Crc was shown to be required for this activity of Hfq but dispensable for Hfq’s role as a sRNA chaperone. Hfq-Crc-mRNA binding leads to repression of translation, particularly of catabolic genes for alternative carbon sources (32), but a surprising finding was that the Hfq-Crc complex also repressed transcription of the virulence-associated transcription factors PtxS, AmrZ and ExsA (31), showing that expression of virulence factors and carbon metabolism may be more closely linked than formerly thought.

Although the methodological innovations presented above are attractive, interpretation of their output data should be done with caution. In NGS, if the sequencing depth is insufficient and/or the signal to noise ratio is too high, low abundance but important interactions can be missed. Moreover, the enrichment steps that rely on retrieval of RNA-BPs or tagged RNA by immunoprecipitation or other types of affinity purification may create non-specific pull-down. These interactions can lead to false positives arising from highly represented but nonspecific pulled-down sequences, or alternatively false negatives for low abundance interactions or RNA complexes with low affinity for their cognate RNA-BP if sequencing depth is too low to detect them. The development of GRIL-Seq (for global small non-coding RNA target identification by ligation and sequencing) using *P. aeruginosa* as model system, provided a significant advance in the field by removing the dependence on RNA-BP interaction and directly enriching ligated sRNA-mRNA complexes, as described in detail below.
In vivo sRNA-mRNA interactions

Due to the limited complementarity between sRNA and mRNA binding pairs, the possibility of bioinformatically predicting all mRNA targets of a given sRNA based on its nucleotide sequence is intrinsically limited, and pre-NGS experimental methods required a priori hypotheses to test. GRIL-Seq is a recent innovation that allows the study of sRNA-mRNA interactions in vivo (33). Developed in P. aeruginosa, this method uses inducible expression of both T4 RNA ligase and a specific sRNA in vivo to create chimaeras of sRNA with their mRNA targets. The total RNA is extracted and enriched for the sRNA of interest before sequencing, revealing all the targets of that sRNA. Thus, the only RNAs that are recovered are ones which were genuinely bound to that sRNA in living cells, removing dependence on a priori predictions and allowing for recovery of low-frequency or transient sRNA-mRNA interactions which might otherwise have been missed. This method led to identification of 17 transcripts targeted by the iron-regulated sRNA PrrF1. Moreover, by combining GRIL-Seq with RNA-Seq in the same experiment, they were also able to differentiate between mRNA targets that were differentially regulated by PrrF1 and RNA targets that were bound but not up- or downregulated, suggestive of alternative modes of action such as “sponge” sRNAs. This approach was further adapted to ligate endogenously expressed sRNAs with their targets without relying on artificial overexpression of the sRNA (34), allowing for identification of the environmental conditions associated with binding of a given sRNA to its targets. Although developed for sRNA target search, it is also possible with this method to use an overexpressed mRNA to identify sRNAs that regulate it. The power of this technique to identify in vivo binding, alternative regulatory mechanisms, and correlation of environmental condition to sRNA-mRNA binding, as well as the possibility to reverse the method by using mRNA instead of sRNA as the target, makes it an excellent candidate to become a key technique for future prokaryotic sRNA research on a large scale.

While the techniques described above have led to many advances in the understanding of post-transcriptional regulation in Pseudomonas, ultimately the phenotypes are caused by protein expression, and while RNA-Seq can predict protein expression levels and subsequent phenotypic traits, it does not formally prove them. Analysis of the Pseudomonas proteome is likely to become an important complement to transcriptomic analysis, not only to validate it but also to define the role of post-translational modifications. Indeed, Pseudomonas sp. post-translationally modify many classes of enzymes and structural proteins in many different ways (reviewed in Gaviard et al. (35)) and it is probable that many of these will be significant for the general physiology of the bacterium and/or for virulence. Proteomic analysis is not only likely to validate predictions generated from transcriptome data, but also to answer questions that NGS approaches cannot currently address.
Fishing for DNA with a transcription factor hook

While RNA-seq and its derivatives aim to discover gene regulation and expression, studies utilizing these methods are not able to determine regulation on a pre-transcriptional level where it is mediated by transcription factors. Furthermore, all regulation-based studies require annotated candidate genes that must be discovered through other means. For these purposes, ChIP-Seq and transposon insertion sequencing (TIS) are revolutionary tools (Figure 2). ChIP-Seq was originally developed in eukaryotic organisms but modified versions of the method soon became valuable for analyzing protein-DNA interactions in prokaryotes (6). ChIP-Seq allows simultaneous discovery of hundreds of regulatory binding sites in a single experiment (Figure 2A), in contrast to classical biochemical assays for DNA-protein binding, in which only one interaction can be studied at a time. In the study of Pseudomonas, ChIP-Seq has brought about the discovery of a vast number of novel regulons, as well as novel functions for regulators which were thought to be already well characterized. Antibiotic resistance, biofilm formation and secretion systems are all well-known virulence systems of Pseudomonas for which ChIP-Seq has significantly expanded the state of the art, as we discuss below.

The implication of biofilm in infection settings has been studied in depth and it is widely understood how biofilm increases antibiotic resistance and nutrient retention, but with new data derived from NGS experiments, previous theories and hypotheses have been challenged. The master regulator of flagella and biofilm-related genes, FleQ, was recently found to be also significantly involved in iron homeostasis (36). Moreover, ChIP-Seq showed that FleQ binding sites overlap with those of AmrZ, a global regulator of environmental adaption, and that 39% of the FleQ and AmrZ-bound sites regulated the expression of iron uptake genes, adding to the importance of biofilm in iron-scarce infection environments. This study also found that FleQ and AmrZ repress each other’s expression, potentially leading to a bistable switch mechanism and indicating that these two factors and their respective regulons are intricately linked. ChIP-Seq has also demonstrated novel regulatory pathways linking biofilm and antibiotic resistance. Recently, the transcriptional regulator pair MdrR1 and MdrR2 (previously PA3898 and PA2100) was shown to both control P. aeruginosa biofilm formation through regulation of phenazine biosynthesis and also to directly interact with the promoter region of the mexAB-oprM efflux pump, repressing its transcriptional expression independently of the MexR regulator (37). Not only does P. aeruginosa biofilm block antibiotics with the extracellular matrix, it also controls the transcriptional expression of other antibiotic resistance factors.

In addition to biofilm formation, the upstream quorum sensing pathways have been shown to be part of a higher interactivity network than previously thought. Schultz et al. (18) provided extensive insight into the regulation network of sigma factors in P. aeruginosa, when they found 43% of the genome to be regulated...
by 10 sigma factors by coupling RNA-seq and ChIP-Seq. Very little, but always function-specific, cross-talk was present between the sigma factors, with RpoN being the common factor that interacted with the most other tested sigma factors (notably RpoH, RpoS, FliA, AlgU, SigX, and RpoD) while also attaining the second largest regulon (680 genes), only outnumbered by the house-keeping sigma factor, RpoD. Since then, RpoN has been shown to directly interact with genes associated with quorum sensing and genes encoding the type VI secretion system with the identification of several new direct targets such as lasI, rhlR, rhlI, pqsR, pqsA, hcpA, and hcpB (38). Not only has a link been established between quorum sensing and the type VI secretion system, but also with antibiotic resistance when RpoN was shown to promote survival in the presence of tobramycin (39). Susceptibility and tolerance could be switched between by disrupting the alternative sigma factor RpoS in a ΔrpoN background, revealing that the two sigma factors together were involved in regulation of the Gac/Rsm pathway (involved in quorum sensing in P. aeruginosa) (40) and the GTP pyrophosphokinase (ppGpp synthetase) relA, demonstrating the cross-talk in adaptive genes suggested by Schultz et al. as well as the massive response RpoN is capable of mounting in given situations.

ChIP-Seq has also aided in determining unexpected regulatory mechanisms of particular DNA binding proteins by showing that they directly bind more, or fewer, targets than was previously supposed. QscR, the quorum sensing-gene repressor, had previously been shown to influence transcription of a large number of genes, but ChIP-Seq revealed that this was due to QscR binding to a single operon of genes of unknown function (PA1895 through PA1897) (41), for which the downstream mechanism has yet to be discovered. Conversely, ChIP-Seq was employed in mapping binding sites of the chromosome segregation factor ParB in P. aeruginosa and revealed that it binds to a specific heptanucleotide sequence motif. However, in addition to binding to the parS site close to the chromosome origin as expected, it was also found at many other loci distributed around the chromosome, allowing for P. aeruginosa ParB to be classified as a nucleoid-associated protein in addition to its chromosome segregation function (42).

An early-adopted variant of ChIP-Seq was its combination with RNA-Seq to differentiate, for a given transcription factor, the directly bound targets from the indirectly regulated ones. While earlier studies performed this analysis for single proteins (15, 16, 36), a major recent advance was provided by the work of Huang et al. (17) which incorporated 20 major virulence-related transcription factors of Pseudomonas aeruginosa and identified their direct and indirect targets among the protein-encoding genome, creating a regulatory network map (PAGnet). This work exposed new potential functions of several of these regulators (especially PhoB, FleQ, AlgR, ExsA and GacA) and revealed highly complex crosstalk among virulence-related transcription factors and an unexpectedly extensive level of indirect regulation.
Since its first application, ChIP-Seq has provided valuable insight into the regulation of *P. aeruginosa* and the understanding of regulatory networks and their interconnection is now much greater. However, as with all methods, ChIP-Seq requires validation through other assays and should be complemented by e.g. electrophoretic mobility shift assays (EMSA) or promoter activity assays. The importance of complementing ChIP-Seq is evident from an attempt at determining the regulon and binding motif of the master regulator of quorum sensing signaling, VqsM, in which the sequencing data had originally omitted key target genes (*lasI, exsA* and *nfxB*) hidden by non-specific noise. These target genes were instead identified bioinformatically from the ChIP-Seq determined binding motif, and then validated as experimentally found targets (43). Only with critical thinking and experimental complementation, is ChIP-Seq an accurate tool, but at that point also a powerful one. Since many *Pseudomonas* transcription factors are response regulators that are activated by external factors via sensor kinases, the regulatory information provided by ChIP-Seq, especially in conjunction with RNA-Seq data obtained under the same experimental conditions, provides vital links between environmental stimuli and the resulting patterns of gene expression, and defines precisely how regulatory networks (such as those controlling biofilm, quorum sensing, antibiotic resistance and expression of virulence factors) are connected, sometimes in unexpected ways. Future ChIP-Seq studies in this opportunistic pathogen are likely to reveal further interconnections between these and other gene networks and enable a deeper understanding of the impact of environmental perturbations on the bacterial transcriptome, particularly when these concern association of the pathogen with its host, or treatment of the pathogen with antibiotics, and of the probable impact on global gene transcription if the bacteria undergo adaptations to their environment which involve alteration of transcription factor function.

**From transposon mutant library to gene network**

Networks of transcriptionally regulated genes come together to produce phenotypes for survival and colonization in specific niches. All bacterial strains have a certain set of genes that must be present for the organism to survive known as the essential genome. Genes necessary for other, specialized circumstances can be grouped together in conditionally essential genomes. The essentiality of a gene relates to the fitness of the organism, where deletion of a non-essential gene does not alter the fitness of the organism while deletion of an essential gene leads to a complete loss of fitness. In between are important but non-essential genes, mutations in which produce an attenuated fitness phenotype, and conditionally essential genes that only produce loss of fitness upon mutation under specific circumstances, and possibly different fitness-implications in other conditions.

In molecular microbiology, the creation of individual gene-knockout mutants has aided the functional discovery of many genes as well as their implication in specific phenotypes but the low-throughput nature
of this process necessitates the establishment of hypothetical gene candidates prior to knockout-mutant creation, and consequently the consistent availability of gene annotation information. TIS removes this requirement in a high-throughput screening process of libraries with hundreds of thousands of transposon-mutagenized bacterial cells and permits the discovery of hundreds of phenotype-associated genes in a single experiment (Figure 2B and (8, 44)).

Gene essentiality is not a novel concept, and multiple databases exist with collections of essential genes from many organisms, most notably the Database of Essential Genes (DEG) (45). The studies included in such databases differ greatly in statistical methods and essentiality cutoff thresholds, with no gold standard. With their development of the statistical method; FiTnEss, coupled with TIS, Poulsen et al. (46) were able to determine a core essential genome of 321 genes for *P. aeruginosa*, accompanied by additional sets of conditionally essential genes associated with the different growth media used in the experiments. The study provides a set of guidelines for determining essential genes including the number of strains and different conditions (media) to test in order to have statistical confidence in the results. However, not only statistical standards must be defined when predicting essential genomes. In a comparison of essential genomes between three studies and their own, Lee et al. (47) found little overlap between the essential genomes, only identifying 141 core essential genes. Though the essentiality of these genes was highly confirmed, several genes generally regarded as essential were absent, such as the replication gene *dnaB* and the RNA polymerase-encoding gene *rpoC*. The list contrasts their own findings of 352 essential genes from six different types of media as well as the 321 genes found by Poulsen et al. testing eight strains in five types of media. Where essential genomes predicted by TIS have previously been established as genes necessary for growth on rich medium, it now seems that “rich medium” is not defined sufficiently stringently, and that essentiality needs to be determined by comparing data from multiple growth conditions taking strain and condition specificity into account.

*The host-pathogen conditionally essential genome*

While the core essential genome provides information on the minimal living conditions of *Pseudomonas*, much research is grounded in determining how, when and why interactions occur with a host. Employing TIS in infection models allow for the determination of fitness-associated genes necessary for colonization and proliferation within a host even at unchanging expression levels. Thus, TIS has been used to identify genes implicated in specific diseases such as cystic fibrosis and wounds, as well as genes responsible for specific virulence systems like type IV pili, biofilm formation and iron acquisition systems. Skurnik et al. (48) performed the first *in vivo Pseudomonas* TIS studies, when they established the conditionally essential genomes of *P. aeruginosa* PA14 colonizing mice ceca and spleens. In a second
publication, the same authors used their previously acquired data to show how disruption mutations in oprD, encoding a carbapenem entry channel, increases PA14 fitness in the murine ceca and spleen (49). A fitness increase associated with bacteria becoming resistant due to the loss of a gene falls into stark contrast with preconceived notions that bacteria must “pay” with fitness advantages to gain antibiotic resistance or other virulence-associated assets. In acute and chronic wounds, TIS has shown that very few infection type-specific genetic pathways exist between the two types of infection, as genes implicated in chemotaxis and flagella are essential only in burn wounds and not in chronic wounds. Additionally, long chain fatty acid metabolism was shown to be essential and highly upregulated in both types of wounds, suggesting that P. aeruginosa wound infections may be treated by interfering with these pathways (50).

Essential genes, especially in a pathogenic species, often make promising targets for treatment of infections, with the ability to eradicate only the pathogenic population while the commensal and mutualistic populations are left unharmed. In many infection settings, the availability of iron is scarce, prompting many human pathogens to incorporate a set of iron acquisition systems. The P. aeruginosa Ferric uptake regulator (Fur) controls both metabolism and virulence and was thought to be essential due to the difficulty in creating a fur deletion mutant. However, with TIS, Fur was shown to only be essential in colony formation and not in planktonic, biofilm or insect model infectious growth, unless grown under iron-rich conditions. Surprisingly, the reason for reduced growth in iron-rich conditions was shown to not be because of a high iron uptake, but rather, due to the dysregulation of the biosynthesis of pyochelin, poisoning the cells in the presence of intracellular iron. Thus, Fur was shown to be a poor target choice for P. aeruginosa infection treatment, despite its initial classification as an “essential” gene and its role in virulence regulation, which would a priori have suggested the opposite (51).

Regardless of disease type, the biofilm-forming capabilities of P. aeruginosa is one of its most well-known traits, with critical implications in infections on both biotic and abiotic surfaces. Crucial for P. aeruginosa biofilm formation and expansion is the characteristic twitching motility mediated by the type IV pili. To generate a whole-genome view of the entire twitching motility gene set, TIS was used to compare twitching-motile to non-motile members of a transposon mutant library by harvesting populations from the outer edge of a twitching motility-elaborated biofilm and comparing it to the center. By this method, the full twitching motility gene set was increased from 44 to 942 putatively implicated genes (52). Of these 942 candidate genes, 42 had previously been known to be involved in twitching motility, while 5 novel genes were experimentally validated in their association with type IV pili; prlC, PA14_66580, pfpl, fliG and motY. The remainder encoded a wide variety of proteins, previously not expected to be implicated in motility,
adding hundreds more to the list of potential type IV pili-associated genes that could serve as potential anti-virulence or biofilm-interrupting drug targets.

Because of its young age, TIS is under constant development with many methodological variants increasing the specificity and simplicity with which a scientific question can be answered. A recent addition is RB-TnSeq, which utilizes unique barcoded transposons to map transposon insertions to mutant genotypes (53). The method was applied by Cole et al. (54) who identified 115 genes required for *P. simiae* colonization of *A. thaliana*. Using RB-TnSeq, they were able to compare the functional growth contributions of genes in 90 distinct *in vitro* conditions. The use of barcodes on transposons removes the need for establishing an input library at every individual experiment, as the barcode-insertion relations in each library can be categorized in a single TIS workflow for use in all further experiments. This ability to multiplex experiments increases the already high throughput of TIS, allowing researchers to easily perform hundreds of distinct mutant library experiments simultaneously. Barcodes may also allow for intricate polymicrobial TIS experiments as mixed species can be separated based on their insertion-barcode profile. A possibility that previously required separate transposons or separate experiments for each transposon mutagenized species (55, 56).

Even more information is made available when sequencing methods are combined for interpretation of multiple types of data. When assaying transcriptomic data, highly expressed genes may seem necessary to the cell fitness and survival, and the combination of TIS and RNA-seq data likely brings forth a hypothesis that gene essentiality and transcriptional expression is negatively correlated, with highly expressed genes attaining few transposon insertions. However, this seems not to be the case. When comparing *P. aeruginosa* grown in defined media to acute and chronic wound infections, no general correlation was found between expression and fitness. Only when subgrouping based on orthologues (Clusters of Orthologous Groups of proteins, COG, (57)) could correlation be seen, and only in the group of metabolically associated genes. In this case, the hypothesis held true; upregulation of metabolically associated genes in the infection model was predictive of fitness defects in mutants, producing a negative correlation (50). Though this is the case for *P. aeruginosa*, it may not be the case for other, more specialized, pathogens. The opportunistic nature of *P. aeruginosa* has not selected for regulation necessary for growth in the “wound” niche and, as such, it would likely require multiple generations for the species to adapt and upregulate genes, a situation found in *P. aeruginosa* infections in the cystic fibrosis lung, where global gene expression adaptation over time has been observed (58).

But attempting to correlate gene expression to gene fitness may not be adequate to see patterns in data. Jensen et al. (59) showed that both in *P. aeruginosa* and in *Streptococcus pneumoniae*, correlations between these two genetic traits are absent, and while genes in specific metabolic pathways could be both
upregulated and conferring a fitness increase, very rarely did genes show both. However, using a mathematical model to increase the scope to include metabolic neighbors (i.e. genes encoding metabolic enzymes that are separated by zero, one or two metabolites in a pathway) revealed that closer neighbors had higher products of fitness and expression changes ($\Delta$fitness $\times \Delta$expression), suggesting that genes closely related in a pathway collectively are correlated in fitness and expression, though individual genes are not. The same analysis from datasets studying antibiotic resistance, indicated that the response to antibiotics is not ordered and neighbors in pathways are not either highly transcribed nor important for fitness. A meta-analysis with all publicly available $S.~pneumoniae$ and $P.~aeruginosa$ datasets, additionally revealed that genes with low expression plasticity (i.e. little change in expression as a function of conditions) also tended to be either essential or conferring a fitness defect, when mutated, in both species (59). These conclusions suggest that first of all, the previous negative correlation hypothesis is likely true, only in situations where modelling is used to describe regulatory relationships within the network and also the genetic interactions between genes, and secondly, that the nature of the stressor determines if transcriptional profiles can be used to determine phenotypically important genes.

Studies of antimicrobial resistance have previously relied heavily on transcriptomics to predict functional importance, but with the poor correlation between the two, a trait that seems universal among bacteria, it is necessary to take precautions when using transcriptomics to predict gene function importance. Basing networks only on gene expression data would exclude many fitness-conferring genes, though, on the other hand, networks based solely on fitness profiles would miss genes that undergo large expression changes. Thus, applying network topological analyses to contextualize high-throughput sequencing experiments has the potential to provide value in predicting and re-evaluating current drug target candidates, consistently made simpler as NGS bioinformatical analysis evolves.

**Evolution of NGS computational analysis; from genome alignments to machine learning-led predictions of antimicrobial resistance**

The clinical relevance of $Pseudomonas~aeruginosa$ made it an early candidate for whole genome sequencing when this field was in its infancy and the first sequence, that of the laboratory strain PAO1, was released in 2000 (60). As more strains were sequenced and made available online, the $Pseudomonas$ research community collectively adopted the strategy of collating all the available genomes in a central and publicly available online database (61) which members of the community kept continuously updated as new gene annotations were added based on published literature and greater functionality was added to the database structure (62). The increasing size of the genome database allowed for greater degrees of confidence in detection of mutations, particularly from clinical isolates, that could be linked to
antimicrobial resistance. However, it soon became clear that such clear-cut genotype-phenotype associations were in the minority, and that to fully define the relationships between genes and antimicrobial resistance (and other) phenotypes on a global scale would require more data, in particular on transcriptome-wide gene expression in as many strains as possible, and from mutation experiments where the requirement of a given gene for a phenotype could be tested empirically.

This data, from RNA-Seq and TIS experiments respectively, has recently allowed much more integrative approaches to be used to study antimicrobial resistance determinants in *Pseudomonas* and to evaluate to what extent expression data (RNA-Seq), gene-associated fitness data (TIS) or a combination of the two could be predictive for antimicrobial resistance phenotypes. To allow three-way comparisons between genome sequences, RNA expression and phenotypes, the database BACTOME was recently developed (63). This database incorporates genome sequences, RNA-Seq datasets and phenotypic information from a panel of *Pseudomonas aeruginosa* clinical isolates (99 at the time of this publication). The phenotypic information in this database focuses on infection-relevant phenotypes including biofilm, *in vivo* virulence in a wax moth larvae model, and resistance levels to a panel of the 5 most common clinically used antibiotics against *Pseudomonas* infection. Hence, associations of nucleotide polymorphisms and/or gene expression levels with a given phenotype could be detected in a centralized manner without individual researchers having to comb the published literature and sequence repositories for data from individual studies, and subsequently investigated for their relevance to clinical outcomes for patients.

Prior to the development of TIS, *Pseudomonas aeruginosa* had been used as the model organism for creation of one of the first ever bacterial arrayed transposon mutant libraries (64), which was then widely shared among the *Pseudomonas* research community. This was used for identifying genes relevant to tobramycin resistance by screening the individual clones of the library and locating the mutants with increased tobramycin sensitivity compared to the wild type (65). The resulting set of 14 genes provided a baseline for the same group to validate the use of TIS to investigate the same phenotype, both to confirm that the same gene set could be identified by this method and to see if it could detect any other genes required for growth in tobramycin (66). Their TIS analysis identified 117 genes associated with increased sensitivity to tobramycin, including 13 of the 14 genes previously found by screening the arrayed library. 85 of these 117 genes were experimentally validated, which confirmed that mutation of genes identified by TIS as necessary for fitness for growth in tobramycin led to at least 2-fold (74 genes) and in some cases 4-fold decreases in the tobramycin MICs of the mutants. The set of 28 genes with 4-fold decreases in MIC contained all the genes identified by the arrayed library screen, showing that TIS analysis has a lower threshold of detection for relevant genes for this antimicrobial resistance phenotype than the arrayed
library screen and is capable of detecting more subtle effects, while the arrayed library screen had only
captured the strongest drug resistance phenotypes. However, TIS proved not to have much predictive
power for the relative importance of individual genes for survival in tobramycin. Some of the genes most
strongly selected against in the TIS analysis did not have correspondingly large decreases in MIC of their
individual mutants, suggesting that the selection pressure in the context of an antibiotic-exposed pooled
library operates differently than in the context of a pure culture. If TIS was to be used for predicting
genotype-phenotype associations regarding antimicrobial resistance, additional data gained from
alternative experimental methods would be required.

Subsequent work sought to rectify this by generating parallel TIS and RNA-Seq datasets from Pseudomonas
cultures exposed to sub-MIC levels of 14 antimicrobial compounds, to examine whether any correlation
could be observed between fitness associations and gene expression level for genes that were associated
with antimicrobial resistance, either for individual compounds or by antibiotic class or mechanism (67).
Similarly to the studies comparing TIS and RNA-Seq in wound infection models (50), the expected inverse
correlation between RNA expression levels and transposon insertion frequency in resistance-associated
genes was not observed; genes that were highly expressed upon antibiotic exposure were not necessarily
very important for fitness in the presence of that antibiotic. However, although gene expression levels were
not useful for predicting which genes would be TIS fitness determinants, orthogonal comparison of RNA-
Seq and TIS datasets across different classes of antibiotics was successful in predicting biologically relevant
cases of antagonism. If a particular antibiotic upregulated expression of a gene that was found to be a
fitness determinant for a second antibiotic, then exposure to the first antibiotic ought to cause increased
resistance during subsequent treatment with the second antibiotic, and this was experimentally validated
for one case (polymyxin B caused increased resistance to gentamicin, tobramycin, neomycin and
ciprofloxacin that was dependent on polymyxin B-induced upregulation of the $mexXY$ genes) (67). Since
treatment of $Pseudomonas$ infections often involves successive courses of antibiotics of different classes,
with the rationale that if resistance to the current antibiotic has evolved during the course of treatment,
the resulting resistant mutants ought to be eliminated by treatment with a different drug, this finding has
significant implications for using this type of data to inform treatment choices so as to avoid such
antagonisms that could not have been predicted otherwise. Conversely, it should be possible to predict
eventual synergies from filtering the data for antibiotics that down-regulate genes which are then fitness
determinants for different drugs, which could then indicate treatment choices with the best chance of
success.
While RNA-Seq profiles of single strains did not show much correlation with antimicrobial resistance phenotypes (59, 67), other studies sought to increase the predictive power of transcriptome analysis by investigating global gene expression profiles across many (clinical) strains for which the antimicrobial resistance phenotypes were known. Combination of data from multiple isolates could then allow identification of subtle transcriptomic effects which would be below the threshold of detection in a single strain but would occur at a statistically significant level if reproduced in a sufficient number of strains sharing the same antimicrobial resistance phenotype. Khaledi et al. (68) employed this approach to examine 135 clinical isolates of *Pseudomonas aeruginosa* for which the MICs of a small set of antibiotics with different cellular targets (ciprofloxacin, ceftazidime, meropenem and tobramycin) had been previously determined. The 135 RNA-Seq profiles (performed in the absence of the antibiotics) were used for transcriptome wide association analysis to identify single genes that displayed a consistent association across the whole panel of strains between expression level and antimicrobial resistance. This revealed a small number of relatively weak gene expression-phenotype associations. However, when the datasets were fed into a machine learning algorithm in order to identify patterns of gene expression from multiple genetic loci, rather than single genes, that would act as markers predicting antimicrobial resistance phenotypes, the algorithm was able to discriminate ciprofloxacin resistant from sensitive strains based on global transcriptome profile even though no single genes had been identified as distinguishing markers for this phenotype and the direct cause of ciprofloxacin resistance in the majority of the resistant strains was a single nucleotide polymorphism in the DNA gyrase gene. Therefore, although the gene expression profiles were a consequence, rather than a cause, of the antimicrobial resistance, they were nonetheless predictive of an isolate’s resistance status provided that enough data was supplied to the algorithm to allow it to detect patterns that discriminated between sensitive and resistant phenotypes.

As the quantity of available bacterial NGS datasets has increased, so too has the complexity of the network connecting the different parameters of bacterial strains, genome sequences, phenotypes, transcription factor regulons, transcriptomes and genome wide fitness data. To take full advantage of the potential of this information, advanced machine learning techniques have recently been developed with the goal of creating an unsupervised machine learning model that is capable of inferring biological information from such data networks without being supplied with *a priori* information about them (here, the *a priori* information could be annotated gene functions, or the antimicrobial resistance/sensitivity status of an organism under investigation, such as in (68)). The key difference between “supervised” and “unsupervised” machine learning models is that a supervised algorithm is trained on existing datasets where both input and output data are supplied (as used in Cornforth *et al.* for discriminating transcriptomes of human infection isolates from those of laboratory-cultured bacteria, (21)), before it is
used for predicting the expected output of novel input data. Conversely, an unsupervised algorithm is allowed to make its own output predictions by building a mathematical model based only on the patterns or clusters it identifies in the input data, without applying any labels or definitions to those patterns or clusters first. Therefore, a supervised algorithm could be expected to be less error-prone (at least in the first iterations) than an unsupervised one, but may be more susceptible to unwitting bias present in the training dataset and/or less good at handling noisy data. Two recent machine learning studies have employed all the publicly available *Pseudomonas aeruginosa* gene expression data acquired with the Affymetrix GeneChip microarray platform to generate unsupervised neural network algorithms and assessed their performance at extracting meaningful biological information from the collected datasets. Whether using a single neural network model (69) or an ensemble network generated from 100 individual models (70), the algorithms were able to make relevant predictions linking gene expression patterns to biological output information, which were experimentally validated for the patterns associated with strain identity, transcriptional response to low oxygen and which medium the bacteria had been grown in. Computational models such as these are likely to provide significant advances to the *Pseudomonas* research community’s ability to interpret large numbers of highly interconnected datasets (Figure 3) and obtain novel biological insights from them.

**Future perspectives and concluding remarks**

While the genomic and transcriptomic studies outlined here have advanced the field of *Pseudomonas* research by allowing systems-level characterisation of the genetic regulatory networks which permit its extreme adaptability to varied environmental conditions, it is clear that the full potential of these methods has not yet been reached. The current standard practice is to prepare transcriptomic or genomic DNA libraries from a population of cells grown in pure culture, but there is much to be gained from “zooming in” to the level of a single cell, or conversely “zooming out” to study mixed cultures containing more than one species of micro-organism. Bacterial single-cell transcriptomics, for which methods are under active development (71, 72) will allow exploration of gene expression patterns associated with micro-level differences in location within an infected tissue to a higher resolution than currently possible, or alternatively to explore phenotypic heterogeneity within a population, in particular aspects of phenotypic heterogeneity that are relevant to chronic infections, such as persister cells (73), response to oxygen gradients within a biofilm (74), heterogeneous transcription factor activity patterns (37) or antibiotic tolerance (75). Genomics or transcriptomics performed on multi-species cultures, for example, mixed-species biofilms or mixed cultures comprising species that are commonly found together in *Pseudomonas*-dominated infection sites, will be able to inform not only about the gene expression patterns and/or
essential genomes of the individual pathogens but also about the polymicrobial interactions involved in this mode of growth and conditional essential genotypes required for proliferation in the presence of other species. The stringent technical challenges of single-cell analysis and mixed-culture experiments are likely to be overcome by the adaptation of the barcoding approach that was spearheaded by TIS experiments (53). Adding barcodes to the NGS adaptors allows detection and compensation for PCR bias in single cell experiments during the cDNA amplification step, since PCR products which are favoured (or not) during the amplification will be reflected in an increased or decreased level of their barcode, which can then be taken into account during data normalisation. Barcoding of mixed-culture TIS experiments will similarly ease the technical bottlenecks of repeated sequencing of output libraries that would otherwise be required, since the unique barcodes identify not only the genetic locus but the strain background from which that gene came, and the range of available barcodes can be expanded easily to fit even larger experiments by simply making the barcodes a base or two longer. Taken together, these approaches are likely to lead to the development of a new generation of biomarkers based on global pathogen gene expression and fitness profiles rather than single gene presence/absence. These could eventually lead to fine-tuning of medical diagnoses and treatments on a case-by-case basis and contribute to the development of “personalised medicine”, especially with the potential for rapid characterisation of the gene expression patterns of a given pathogen interacting with a patient’s own cells, as the speed of the technology increases. At the level of hardware, as NGS sequencers are simplified and miniaturised, it can be anticipated that they will become indispensable tools for field as well as laboratory research, for example to track geographical spread and/or genetic evolution of outbreaks of infectious disease.

While the integration of “omics” data into medical practice presents a valuable opportunity for better targeting of patient treatment, the challenges are also evident. For this approach to succeed, two main criteria must be fulfilled; the sequencing data must provide a biologically relevant insight that is directly translatable to patient care, and this biologically relevant insight must be obtained quickly enough that it can inform treatment choice for the patient at the appropriate time. Sample processing methods and sequencing software and hardware developments are quickly overcoming the challenges of small bacterial sample size, the relative over-representation of human relative to bacterial transcripts (if a dual RNA-Seq approach is being used) and speed of obtaining and analysing the data. However, in order to obtain clinically useful biological information, it is essential that the right questions are asked and the right technique is used to answer them. It may be the case that a question about, for example, the antimicrobial resistance status of a bacterial pathogen is one that cannot be appropriately answered by examining its genome or transcriptome. Indeed, several of the works highlighted in this Review (50, 59, 67) suggest that it is not, and that even a combination of global gene fitness profiling by TIS and transcriptome sequencing
by RNA-Seq is not sufficient to draw clinically relevant conclusions. To counteract these challenges, the addition of proteomics analysis to sequencing approaches may contribute towards filling the gaps that transcriptome or gene fitness data cannot fill. Mass spectrometry techniques are being developed to assess and predict antimicrobial resistance status of pathogens at the proteome level, and although these are currently not sufficient for clinical application to organisms with complex multifactorial antimicrobial resistance mechanisms such as *Pseudomonas aeruginosa* (76) they are likely to improve in accuracy as the technology improves and more whole-proteome information (especially from clinical isolates) becomes available. Integration of such data, together with existing genomic, RNA-Seq, ChIP-Seq and phenotypic data sets into networks or databases spanning large numbers of isolates could eventually permit more accurate predictions to be made by the use of association studies, which would be expected to possess much more statistical power than NGS analyses performed on one strain alone. In order to interrogate data networks of this complexity, however, it is likely that machine learning algorithms will become as indispensable and mainstream in the near future as the BLAST algorithm (77) is today.

With more and more research in NGS fields, more caveats are also brought to light, and though it has long been clear that high-throughput experiments always require validation, and that NGS experiments complement each other well, it is a recent discovery that some studies may actually require multiple types of high-throughput experiments in order to draw reliable conclusions from them. In this review, we have presented insight into multiple papers, showing how gene expression and fitness do not generally correlate, which alters the notion that transcriptomic studies are able to predict genes important for fitness. Rather, the experimental condition is determining for knowing if expression is indicative of functional importance. In the future, studies (especially those focusing on developing new pathogen treatment targets) may have to investigate multiple types of high-throughput data, validating their findings in each with the findings in the other. Moreover, new mathematical models, likely derived from ChIP-Seq and RNA-Seq data, may be required to fully define topological network relationships between genes that are involved in virulence, quorum sensing, biofilm or antimicrobial resistance, analogous to those that have been constructed for metabolic networks. It is conceivable that a greater degree of coherence between RNA-Seq and TIS data may be observed when examined in the framework of a context-specific *in silico* model than examining data derived from antibiotic perturbation in the context of a model designed to predict metabolic perturbation effects (59). Combined with computational power, these mixed studies may determine fitness gene, regulatory factor and expression levels across multiple conditions in a single study, accelerating the search for treatment targets of *P. aeruginosa* and other pathogens.
Acknowledgements

This work was supported by grants from the Swiss National Science Foundation, Marie Heim-Vögtlin scheme (PMPDP3_158295), the Fondation Pierre Mercier pour la Science and intramural funding from the University of Southern Denmark (to CLK). We thank Sergi Torres Puig for critical reading of the manuscript. We also recognize the contributions of the other relevant studies in this field that could not be cited here due to space constraints.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Figure legends

Figure 1. Development of NGS approaches to study RNA-BP and sRNA interactions with the global transcriptome. How the basic RNA-Seq method (centre) has been developed into more advanced techniques (inner ring), and the biological information that is gained from them (outer ring). Clockwise from top-left: Dual RNA-Seq studies compare transcriptional changes in 2 species (e.g during host-pathogen interaction) by extracting total RNA and mapping the reads to both the host and the pathogen genomes. MAP-Seq & GRIL-Seq study sRNA interaction: MAP-Seq employs a MS2 RNA affinity tag bound to a sRNA of interest which will be co-purified with its interacting RNA and sequenced. GRIL-Seq involves co-expression in vivo of a sRNA of interest and T4 RNA ligase creating chimaeras; total RNA is enriched for transcripts containing the sRNA chimaeras and sequenced. RIP-Seq & CLIP-Seq identify RNA-BP (shown here as Hfq, but applicable to other RNA-BPs) targets: RIP-Seq by immunoprecipitation of RNA targets via pull down of the RNA-BP of interest in cell lysate followed by RNA-Seq. CLIP-Seq by in vivo UV exposure to covalently cross-link RNA to proteins before co-purification of the RNA-protein complexes and RNA-Seq. RNA trimming by ribonucleases outside of the binding region enables mapping of the binding region at single-nucleotide resolution. RIL-Seq uses immunoprecipitation as in RIP-Seq and crosslinking as in CLIP-Seq to bind RNAs to the RNA-BP. After enzymatic digestion, RNAs are ligated and subjected to RNA-Seq. RNA–RNA interactions are revealed after mapping and identification of chimeric reads. CLASH uses the same method as RIL-Seq but the bound RNA is trimmed using an RNase and RNA linkers are ligated to the immobilized RNA molecules in the RNA–BP.
complexes. Finally, coupled RNA molecules are ligated into one single molecule of two different types, either single or chimaeric fragments. dRNA-Seq & pRNA-Seq reveal respectively transcription start sites and RNA processing sites. dRNA-Seq discriminates primary from processed 5’ends by sequencing a differential cDNA library. One is prepared from untreated total bacterial RNA and the other is enriched for primary transcripts by terminator exonuclease treatment that degrades 5’P but not 5’PPP RNA. pRNA-Seq sequences specifically processed RNA though adaptors ligated in vitro on 5’P and 3’OH ends, followed by reverse transcription, amplification of fragments with adaptors by PCR and sequencing.

Figure 2. Application of NGS to achieve molecular resolution of DNA-protein interactions, core genomes and conditional-essential genes.

(A) Chromatin immunoprecipitation sequencing works by cross-linking proteins to DNA with formaldehyde in actively growing cells. The DNA is purified, fragmented, and DNA-protein complexes are enriched for in an immunoprecipitation assay using an antibody specific to the protein of interest, (discarding the un-enriched DNA-fragments). The crosslinks are reversed, and a DNA library is prepared from the isolated fragments. The immunoprecipitant DNA library is sequenced alongside an un-enriched, non-treated sample representing the background signal. Protein binding sites are found as regions of the genome with enrichment in read count in the immunoprecipitated sample compared to the background sample.

(B) Transposon insertion sequencing studies start with the creation of a saturated and complex transposon mutant library. The library is grown under input and output test conditions, input conditions being the neutral baseline to which the results of the desired output condition are compared. Viable mutants from the experiments are then harvested and genomic DNA is extracted and prepared for next-generation sequencing. The samples are sequenced, and data analysis comparing the transposon insertion distribution between input and output to identify essential (gene C) and conditionally essential genes (gene A) is performed.

Figure 3. Multi-dimensional interconnectivity between “omics” data sets requires highly efficient computational tools.

With the increasing quantity of available bacterial NGS datasets, follows a higher degree of complexity in the network connecting genome sequences, phenotypes, transcription factor regulons, transcriptomes and genome wide fitness data between strains. With the struggle to understand the complexity of the omics network in a single strain, it has become clear that the full potential of this information is only available as more advanced computational tools, such as machine learning and neural networks, are developed and integrated to fit specific research purposes.
References


