International Journal of Medical Microbiology 305 (2015) 339–347

Contents lists available at ScienceDirect

International Journal of Medical Microbiology

journal homepage: www.elsevier.com/locate/ijmm

Choose wisely: Network, ontology and annotation resources for the analysis of *Staphylococcus aureus* omics data

J.A. Broadbent[a,c,*], D.L. Sampson[a,c], D.A. Broszczak[a,c], Z. Upton[a,c], F. Huygens[b,c]

[a] Injury Prevention and Trauma Management, Queensland University of Technology, Brisbane, QLD, Australia
[b] Chronic Disease and Aging, Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, QLD, Australia
[c] School of Biomedical Sciences, Faculty of Health, Queensland University of Technology, Brisbane, QLD, Australia

ARTICLE INFO

Article history:
Received 25 July 2014
Received in revised form 21 January 2015
Accepted 9 February 2015

Keywords:
Staphylococcus aureus
MRSA
Omics
Molecular network
Gene ontology
Gene ontology annotation
Systems biology
Non-model organism
Bioinformatics

ABSTRACT

Staphylococcus aureus (*S. aureus*) is a prominent human and livestock pathogen investigated widely using omic technologies. Critically, due to availability, low visibility or scattered resources, robust network and statistical contextualisation of the resulting data is generally under-represented. Here, we present novel meta-analyses of freely-accessible molecular network and gene ontology annotation information resources for *S. aureus* omics data interpretation. Furthermore, through the application of the gene ontology annotation resources we demonstrate their value and ability (or lack-there-of) to summarise and statistically interpret the emergent properties of gene expression and protein abundance changes using publically available data. This analysis provides simple metrics for network selection and demonstrates the availability and impact that gene ontology annotation selection can have on the contextualisation of bacterial omics data.

Introduction

The rapid evolution of *Staphylococcus aureus* (*S. aureus*) combined with over-use/misuse of antibiotics has seen this organism transition from a treatable inconvenience to a major threat to global health and economic security (Editorial, 2013). Infections caused by this organism range from mild to moderate skin and soft tissue infections to severe invasive infections of the bone, lungs and heart. These infections lead to mortality in 20–50% of cases (Miro et al., 2005) and underpin a US$86 billion burden on American patients (Filice et al., 2010; Klein et al., 2013) and $830 million–$9.7 billion on the American healthcare system per annum (Klein et al., 2007). Whilst once confined to healthcare settings, drug-resistant *S. aureus* is now widespread in the community (Nimmo et al., 2006, 2013). In addition to drug resistance, these community-associated strains display signs of enhanced virulence when compared to healthcare-associate strains (Gordon and Lowy, 2008; Otto, 2010). Given this phenomenon and the ever-increasing health and financial burden of this organism, *S. aureus* has been the subject of numerous investigations that seek to understand its biochemistry (Basell et al., 2014; Hessling et al., 2013; Michalik et al., 2012), interpret the mode action of antimicrobial agents (Overton et al., 2011), or identify new antimicrobial targets (Cherkasov et al., 2011). As *S. aureus* elicits disease through a combination of pathogenic, virulence and antibiotic resistance factors – traits predetermined by a genetic program and executed by metabolic and functional biochemistry – this organism has been widely investigated using omics technologies.

Proteomics and transcriptomics are cornerstone technologies in systems biology, allowing the investigation of function through complex biomolecular interactions and dynamics. The vast amount of data produced through these approaches necessitates the ability to contextualise broad changes in gene expression or protein abundance. Such contextualisation was for some time limited to the assignment of molecules to specific metabolic pathways or molecular classes. However, more recent advances have integrated robust statistics (Huang et al., 2009), ontology resources (Ashburner et al., 2000; Ogata et al., 1999) and molecular networks (Overton et al., 2011; Solis et al., 2014) to provide a more comprehensive interpretation of global molecular perturbations. These resource types have for some time represented a core toolkit for discipline-standard interpretation of omics data obtained from model organism investigations. While their uptake in microbiological investigations (and

* Corresponding author at: Institute of Health and Biomedical Innovation, Queensland University of Technology, Kelvin Grove Campus, 60 Musk Avenue, Brisbane 4059, QLD, Australia. Tel.: +61 7 3138 6201; fax: +61 7 3138 6030.
E-mail address: j2.broadbent@qut.edu.au (J.A. Broadbent).

http://dx.doi.org/10.1016/j.ijmm.2015.02.001
1438-4221/

non-model organisms more generally) is not yet routine, robust statistical and network contextualisation of *S. aureus* omics data have recently been reported in the literature (Conlon et al., 2013; Marbach et al., 2012; Overton et al., 2011; Solis et al., 2014). These investigations highlight both popular and novel resources that may enhance omics investigations. However, the presence of analogous *S. aureus* information resources, originating from various unique sources, calls into question the quality and value that each provides in terms of its ability to robustly contextualise data. This limits the ability of researchers to select the resources that can meaningfully and efficiently exploit the results of their omics experiments.

Here we have amassed novel and previously reported network and ontology resources used for the contextualisation of *S. aureus* omics data. Through meta-analyses and performance evaluation using publically available micro-array and proteome data sets, we have quantified the ability of these resources to interpret broad changes in gene expression and protein abundance. This analysis forms a guide to the optimal interpretation of *S. aureus* data based on the empirical measurement of performance. The use of these optimal resources can be expected to enhance data exploitation and thereby lead to new discoveries and improved hypothesis formation and experimental design. Such advancements will lead to a greater understanding of this important pathogen thereby providing necessary information for improved disease prevention, diagnosis, treatment and management. Indeed, the application of robust information and statistical resources will enable more accurate and reliable data generation that will enhance our understanding of the pathogenic and antibiotic resistance mechanisms that *S. aureus* utilizes in causing severe and life threating human disease.

Materials and methods

Network graph collection, unification and analysis

Network graphs were acquired from literature and database sources. Literature-sourced networks included the Cherkasov experimental protein–protein interaction network (Cherkasov et al., 2011), Marbach inferred gene regulatory network (Marbach et al., 2012) and the Overton inferred protein–protein interaction network (Overton et al., 2011). Database-sourced networks were acquired from String v9.05 (String-db.org), the Kyoto Encyclopaedia of Genes and Genomes (KEGG; http://www.genome.jp/kegg/), Virulence Factor Database (VFDB; http://www.mgc.ac.cn/VFs/; release 3) and RegPrecise v3.1 (regprecise.lbl.gov/RegPrecise/). The Marbach and String networks were obtained in formats compatible with Cytoscape import and downstream analysis, while the remaining networks required some level of construction and/or ordered locus name (OLN) unification.

De novo construction of networks was performed for the KEGG, VFDB and RegPrecise datasets. The KEGG and VFDB networks were constructed by associating OLNs and their respective classification terms obtained from the KEGG BRITE functional hierarchy and virulence factor classification, respectively. RegPrecise gene regulatory network data were available only for the related strain N315. Consequently, N315 OLNs were mapped to MU50 OLNs using the homology search at the J. Craig Venter Institute – Comprehensive Microbial Resource (JCVI-CMR). The new OLNs were then used to construct a network connecting MU50 OLNs based on the RegPrecise gene regulatory networks for *S. aureus*.

Additional network unification was required for the Cherkasov and Overton networks, which were both originally constructed using the related strain MRSA-252 (Cherkasov et al., 2011; Overton et al., 2011). For the Cherkasov protein–protein interaction network, RefSeq identifiers were mapped to OLNs using the ID Mapping tool at the UniProtKB (http://www.uniprot.org/uniprot/). The resulting MU50 OLNs were then obtained using the homology search at the JCVI-CMR. The Overton network MRSA-252 OLNs were mapped to MU50 OLNs at the JCVI-CMR. Following unification, networks were assessed for their genome coverage (including the proportion of OLNs not found in other networks), virulence factor coverage and gene regulatory network coverage. All networks were imported into Cytoscape v2.8.2 to visualise network topology and key properties (Shannon et al., 2003). Ontology analysis was performed in Cytoscape using the BiNGO application as described in the following sections.

Detailed network information is provided in supplementary information (Supplementary Table S1).

Information resource collection, unification and meta-analysis

Information resources were acquired, extracted or created from various sources. The gene ontology (GO) was acquired from the Gene Ontology project (http://www.geneontology.org/; 04/04/2013), while the MU50 UniProt gene ontology annotations for cellular component, molecular function and biological process were acquired from the UniProt-GOA (www.ebi.ac.uk/GOA; 04/04/2013) by filtering for *S. aureus* MU50 taxonomy (Tax ID: 158878). The Affymetrix gene ontology annotations for cellular component, molecular function and biological process were extracted from the Affymetrix platform (GPL1339) data table obtained from the National Centre for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) using an in-house *R* script. Gene ontology annotations for cellular component, molecular function and biological process were extracted from the JCVI-CMR using the genome search tool. Annotations against The Institute of Genome Research (TIGR) roles ontology were also extracted from this source. TIGR and KEGG ontologies were then constructed using the TIGR roles and KEGG BRITE functional hierarchy classifications, respectively. Ontology annotations mapped against the GO were quality controlled using CateGOrizer (http://www.animalgenome.org/bioinfo/tools/catego/) to identify obsolete or alternate GO terms. These were then mapped to current or dominant terms, respectively. TIGR role terms with low value (*e.g.* unknown function) were not used in subsequent meta-analysis. Term and gene-term overlap between gene ontology annotations was assessed using Venn Diagram Generator (http://www.pangloss.com/seidel/Protocols/venn.cgi). Quantitative Venn diagrams were created using Google Chart. Summary information regarding gene ontologies and gene ontology annotations are supplied as supplementary information (Supplementary Tables S2 and S3).

Omics data and analysis

Publically-available micro-array data sets were used to evaluate the performance of gene ontology annotation resources. In this regard six experiments were acquired from the NCBI GEO under experiment accession numbers GDS1666, GDS2105, GDS2812, GDS2814, GDS2983 and GDS3136. To facilitate ontology analysis, MU50 strain protein accession numbers were manually filtered from the Affymetrix platform (GPL1339) data table obtained from the GEO. The resulting 2629 open reading frames (ORFs) were then applied to each data file as input gene names for further analysis. Expression data for these 2629 ORFs were then $\log_2$ transformed and analysed using permutation statistics in the SAM (Statistical Analysis of Micro-arrays; Stanford University) v4.00 *R* add-on (Tusher et al., 2001). The parameters used included two class unpaired test, median centre normalisation and 1000 permutations. All other parameters were set to default. The resulting data were then filtered to obtain genes with