211(F)

A Comparative Genomics Approach to Rapid Diagnostic Development

Daniel Fasulo, Ph.D.—Pattern Genomics, LLC (Madison, CT) and Richard Winegar, Ph.D.—MRIGlobal (Palm Bay, FL)

Correspondence: daniel.fasulo@patterngenomics.com **Additional Materials:** http://patterngenomics.com

Abstract

In the event of a disease outbreak, there is a need to rapidly characterize the pathogen and develop broadly-deployable, rapid detection kits for surveillance and diagnosis. We present a comparative genomics approach to finding distinguishing genomic features in a class of pathogens and developing PCR primer sets for molecular identification in a single, efficient analysis. In our approach, a representative set of isolates undergoes genome sequencing and assembly. Software takes as input the genomes, constraints related to oligonucleotide design, and a label for each genome indicating whether the genome represents a positive or negative sample. The output includes data about the relationships among the genomes, regions that distinguish the positive and negative examples, and a panel of oligonucleotide primers and probes that can distinguish the positives from the negatives.

To demonstrate the approach, we developed a compact primer set that distinguishes methicillin-resistant *Staphyloccus aureus* (MRSA) from methicillin-sensitive *Staph. aureus* (MSSA) and coagulase-negative *Staphyloccus* (CoNS) of any resistance status. By analyzing over 110 whole genomes including our own draft sequences of 3 emerging MRSA strains, we were able to derive primers that, in combination, handle challenges such as empty SCCmec cassettes, strains carrying the recently-identified *MecC* resistance gene, and community-acquired strains bearing unusual SCCmec variants. A single primer pair is sufficient to detect approximately 85% of the MRSA genomes, but the remainder represents primarily community-acquired strains with a higher level of genomic diversity.

We present qPCR validation data on reference samples including 10 genetically diverse MRSA strains including recent community-acquired isolates, 3 MSSA strains, 4 CoNS strains including one bearing resistance to methicillin, and human DNA. In an initial round of validation, the primer set was 100% accurate but two samples showed unexpected marker patterns and a third showed evidence of low-level contamination. Next-generation sequencing was used to analyze these samples and indicated that the contaminated sample had a mixture of genomic sources, while the two unusual marker patterns were linked to rare sequence variants of the SCCmec region.

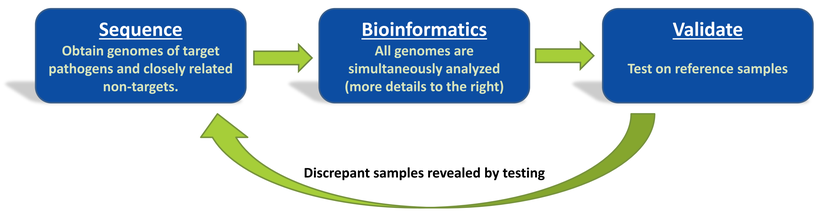

Figure 1. Summary of NGS-enabled, data-driven assay design process.

Introduction

Molecular diagnostics have the potential to provide sensitive, specific, and rapid detection of pathogens. One major component of a molecular diagnostic is the design of oligonucleotide primers and / or probes that are well conserved so that they react to all of the intended targets, while also being specific enough to avoid cross-reactivity with other sources of DNA. The selected sequences must also satisfy various physical constraints determined by the testing chemistry.

Next-generation DNA sequencing enables a rapid, data-driven approach to assay design. The genomes of many isolates are quickly and inexpensively sequenced and assembled into draft genomes. A specialized bioinformatics platform is used to compare the target isolate genomes to the genomes of related but distinct organisms and select primers and probes that specifically identify the target and satisfy the physical constraints. The approach is summarized in Figure 1.

The advantages of this approach are that it is *general*, applying to any situation where isolates can be gathered and sequenced, it is *unbiased*, since entire genomes are scanned without pre-selection of genes or a reference sequence, it is *rapid*, because sequencing and assembly of most viral and bacterial pathogens is routine and the software runs in minutes to hours, and the resulting assays are *robust* as long as the population of samples used in the analysis is representative.

Experimental Setup

In order to validate our methodology on a difficult and relevant assay design problem, we chose to focus on *Staphylococcus aureus* (SA) infections and design a quantitative PCR primer set for the detection of methicillin-resistant SA (MRSA). The primer set was required to detect the presence of MRSA without interference from methicillin-sensitive SA (MSSA), coagulase-negative SA (CoNS) of any resistance status, or other common skin or respiratory pathogens. For simplicity, we chose a standard SYBR Green kit for qPCR product detection rather than a labeled probe. Primers were tested on 19 DNA samples (ATCC), including

- 10 MRSA samples: ATCC Methicillin Resistant Staphylococci Microbial Panel MP-2 (Samples #1-#7), ATCC 700699D-5 (#8), BAA-1556D-5 (#9), and ATCC 43300 (#10).
- 3 MSSA samples: ATCC 6538D-5 (#11), ATCC 25923D-5 (#12), ATCC 10832D-5 (#13).
- 4 other *Staphylococcus* samples: *S. saprophyticus* ATCC 15305D-5 (#14), *S. epidermidis* strains ATCC 12228D-5 (#15), ATCC 35984D-5 (#16, note CoNS-MR), ATCC 14990 (#17).
- 2 unrelated negative controls: *Streptoccocus pyogenes* BAA-1063D-5 (#18), and human cell line HTB133-D (#19).

Bioinformatics Methods

We created a custom bioinformatics platform to facilitate our data-driven assay design approach. Given two sets of genomes, one of which represents the target pathogens and the other representing non-target sequences, the platform attempts to directly answer the question, "What is the difference between these two sets, and how can one test for it?" The genomes need not be finished; rough draft genomes output from a typical genome assembler are sufficient. The general flow of the process is shown in Figure 2 below.

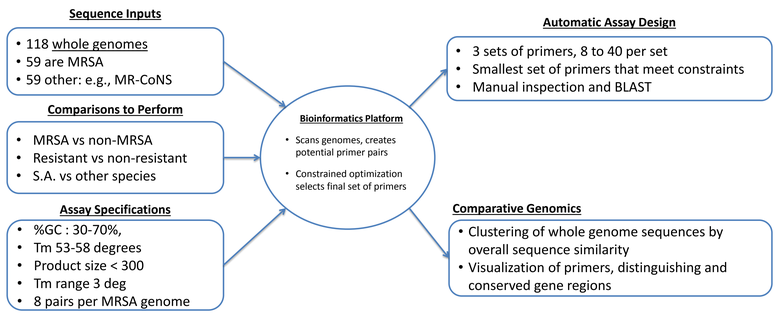

Figure 2. Summary of bioinformatics process, using MRSA study as an example.

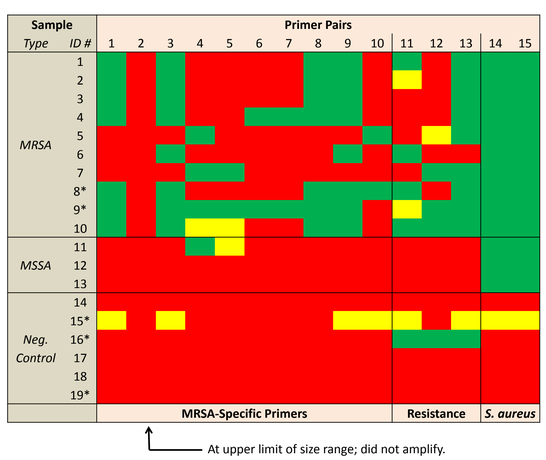

Figure 3. qPCR results of Phase I primers on 19 samples. The rows are samples, the columns are primer pairs, and the entries summarize amplification. Green denotes a C_T value less than 30 indicating a strong signal, Yellow denotes a C_T value between 30 and 35 indicating a weak signal, and Red denotes a C_T value of 35 or higher indicating no signal. A starred sample(*) indicates a genome that had been sequenced prior to the experiment.

← Unexpected positives in primer pairs 4 and 5

← Unexpected positives; contamination?

← Expected; resistant CoNS strain

Primer Set Design and Testing

Design Iteration 1

We used 102 whole genomes to design a candidate set of qPCR primers using the bioinformatics method described previously and publicly-available MRSA data. We designed 3 distinct primer categories:

- Primers that are specific to MRSA, meaning that they do not generate product in non-MRSA genomes
- Primers that separate resistant strains from non-resistant, regardless of species
- Primers that are specific to SA as a species.

The results are depicted with a heat map in Figure 3. The data in the figure shows that a strong positive signal from at least one primer set in each of the 3 categories is sufficient to definitively identify MRSA; however, the results have imperfections as highlighted in the figure.

DNA Sequencing

We performed a single DNA sequencing run on an Illumina MiSeq, sequencing samples #5, #6, #7, #9, #11, and #15. The sample preparation was performed with the Nextera XT kit with an average fragment length 685 base pairs, and the 2x250 paired end protocol was used. All genomes except #15 were assembled *de novo* using Newbler v2.9 from 454. The goals of the sequencing run were to

- Eliminate the false positives in genome #11 and provide alternate primer pairs that detect #7
- Have greater representation of emerging strains in the assay design, using genomes #5, #6, and #7
- Assess the quality of our sequencing and assembly using reference strain #9 (data not shown)
- Prove contamination of sample #15 by deep sequencing (see Figure 4 below)

Design Iteration 2

Using data from the sequencing run described above, a new set of primers detecting MRSA strains specifically was created and testing. The total number of genomes used was 118 after also incorporating newly-available data from GenBank. The final recommended set of primers from the two iterations is in Figure 5 below.

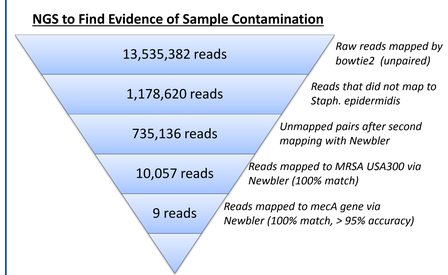

Figure 4. Mapping of reads from sample #15, *S. epidermidis* ATCC-12228, finds a low percentage of unambiguously MRSA DNA. Reads were mapped to the available reference genome for this strain, including plasmids, by two orthogonal methods. Those reads that did not map by either method were then compared to a finished MRSA reference genome.

Sample Type	ID	Primer Pairs 9	17	19	11	13	14
MRSA	1						
	2						
	3						
	4						
	5						
	6						
	7						
	8						
	9						
	10						
MSSA	11						
	12						
	13						
Neg. Control	14						
	16						
	17						
	18						
	19						
		MRSA			Resistance		S. aureus

Figure 5. Final recommended primer sets for MRSA detection derived from both iterations. Primer pair 19 is included due to predicted sensitivity to additional strains *in silico.*

Discussion

MRSA detection is a difficult task, owing to the diversity of SCCmec types and the existence of closely-related CoNS. Nevertheless, when given representatives of many MRSA and non-MRSA genomes, in less than an hour advanced algorithms are able to find patterns of primer pairs in the raw data that are capable of making the distinction. Furthermore, in the case where new samples shown an unexpected amplification results, we show that this methodology easily adapts by incorporating new sequencing data. Hence it is feasible to design a targeted detection assay nearly as quickly as one as able to obtain isolates and sequence them. We believe this general framework will be of great value in understanding, tracking, and diagnosing emerging diseases and outbreaks.