Hindawi Publishing Corporation
EURASIP Journal on Bioinformatics and Systems Biology
Volume 2007, Article ID 23054, 7 pages
doi:10.1155/2007/23054

Research Article

Genome-Wide Analysis of Intergenic Regions of *Mycobacterium tuberculosis* H37Rv Using Affymetrix GeneChips

Li M. Fu[1] and Thomas M. Shinnick[2]

[1] *Pacific Tuberculosis and Cancer Research Organization, 8 Corporate Park, Suite 300, Irvine, CA 92606, USA*
[2] *Centers for Disease Control and Prevention, Atlanta, GA 30333, USA*

Received 24 April 2007; Accepted 14 August 2007

Recommended by Z. Jane Wang

Sequencing the complete genome of *Mycobacterium tuberculosis* H37Rv is a major milestone in the genome project and it sheds new light in our fight with tuberculosis. The genome contains around 4000 genes (protein-coding sequences) in the original genome annotation. A subsequent reannotation of the genome has added 80 more genes. However, we have found that the intergenic regions can exhibit expression signals, as evidenced by microarray hybridization. It is then reasonable to suspect that there are unidentified genes in these regions. We conducted a genome-wide analysis using the Affymetrix GeneChip to explore genes contained in the intergenic sequences of the *M. tuberculosis* H37Rv genome. A working criterion for potential protein-coding genes was based on bioinformatics, consisting of the gene structure, protein coding potential, and presence of ortholog evidence. The bioinformatics criteria in conjunction with transcriptional evidence revealed potential genes with a specific function, such as a DNA-binding protein in the CopG family and a nickle binding GTPase, as well as hypothetical proteins that had not been reported in the H37Rv genome. This study further demonstrated that microarray-based transcriptional evidence would facilitate genome-wide gene finding, and is also the first report concerning intergenic expression in *M. tuberculosis* genome.

1. INTRODUCTION

Unraveling the complete genome sequence of *Mycobacterium tuberculosis* H37Rv [1] has led to a better understanding of the biology and pathogenicity of the organism. This is a major advance in combating tuberculosis (TB), a deadly infectious disease caused by *M. tuberculosis*. With this accomplishment, new molecular targets for diagnostics and therapeutics can be invented at a fast pace by searching the genome.

To utilize the information embedded in a genome, the genome must be annotated thoroughly. In essence, genome annotation is to identify the locations of genes and all of the coding regions in a genome, and determine their protein products as well as functions. As hundreds of bacterial genome sequences are publicly available and the number will soon reach the milestone of 1000, the need for automated, large-scale, high-throughput genome annotation is rapidly increasing [2–4]. A recent study indicates that many genomes could be either over-annotated (too many genes) or under-annotated (too few genes), and a large percentage of genes may have been assigned a wrong start codon [5]. Even if the original genome annotation looks accurate and complete upon submission, it needs to be updated on a regular basis in accordance with new experimental evidence and knowledge that is evolving over time. However, reannotation of the whole genome is not very fruitful, as most of the genes have been identified in the first annotation. For example, the reannotation of the H37Rv genome resulted in about 2% of new protein-coding sequences (CDS) added to the genome.

Some intergenic sequences in *M. tuberculosis* genome exhibit expression signals, as detected by the Affymetrix GeneChip. The same observations have been made for other bacteria, such as *Bacillus subtilis* [6], and also in the eukaryotic system [7]. At present, it is not clear whether or how intergenic expression represents gene activity. Here, we conducted a genome-wide analysis using the Affymetrix GeneChip to explore genes contained in the intergenic sequences of the *M. tuberculosis* H37Rv genome. Potential protein-coding genes were determined based on the bioinformatics criteria [8, 9] consisting of the gene structure,

protein coding potential, and presence of ortholog evidence. We present the first report concerning intergenic expression in *M. tuberculosis* genome and show that microarray-based transcriptional evidence would facilitate genome-wide gene finding.

2. MATERIALS AND METHODS

2.1. Bacterial culture of M. tuberculosis

M. tuberculosis strain H37Rv was obtained from the culture collection of the Mycobacteriology Laboratory Branch, Centers for Disease Control and Prevention at Atlanta, GA, USA. A portion of a recently frozen stock was inoculated into 5 ml of complete Middlebrook 7H9 broth (7H9) supplemented with 10% albumin-dextrose-catalase v/v (Difco Laboratories, Detroit, Mich, USA) and 0.05% Tween 80 v/v (Sigma, St. Louis, Mo, USA) and incubated at 37°C for 5 days. Then the culture was transferred into 50 ml of 7H9 media and incubated at 37°C with 50 rpm shaking until the OD600 reached 0.35. The cells were harvested by centrifugation for RNA preparation.

2.2. RNA isolation

Bacterial lysis and RNA isolation were performed following the procedure of [10] at the CDC lab. (Atlanta). Briefly, cultures were mixed with an equal volume of RNALater^{TM} (Ambion, Austin, Tex) and the bacteria harvested by centrifugation (1 minute, 25 000 g, 8°C) and transferred to Fast Prep tubes (Bio 101, Vista, Calif) containing Trizol (Life Technologies, Gaithersburg, Md). Mycobacteria were mechanically disrupted in a Fast Prep apparatus (Bio 101). The aqueous phase was recovered, treated with Cleanascite (CPG, Lincoln Park, NJ), and extracted with chloroform-isoamyl alcohol (24 : 1 v/v). Nucleic acids were ethanol precipitated. DNAase I (Ambion) treatment to digest contaminating DNA was performed in the presence of Prime RNase inhibitor (5′−3′, Boulder, Colo). The RNA sample was precipitated and washed in ethanol, and redissolved to make a final concentration of 1 mg/ml. The purity of RNA was estimated by the ratio of the readings at 260 nm and 280 nm (A260/A280) in the UV. 20 ul RNA samples were sent to the UCI DNA core and further checked through a quality and quantity test based on electrophoresis before microarray hybridization.

2.3. Microarray hybridization

In this study, we used the antisense Affymetrix *M. tuberculosis* genome array (GeneChip). The probe selection was based on the genome sequence of *M. tuberculosis* H37Rv [1]. Each annotated open reading frame (ORF) or intergenic region (IG) was interrogated with oligonucleotide probe pairs. An IG refers to the region between two consecutive ORFs. The GeneChip represented all 3924 ORFs and 740 intergenic regions of H37Rv. The selection of these IGs in the original design was based on the sequence length. Twenty 25-mer probes were selected within each ORF or IG. These probes are called PM (perfect-match) probes. The sequence of each PM probe is perturbed with a single substitution at the middle base. They are called MM (mismatch) probes. A PM probe and its respective MM probe constitute a probe pair. The MM probe serves as a negative control for the PM probe in hybridization.

Microarray hybridization followed the Affymetrix protocol. In brief, the assay utilized reverse transcriptase and random hexamer primers to produce DNA complementary to the RNA. The cDNA products were then fragmented by DNAase and labeled with terminal transferase and biotinylated GeneChip DNA Labeling Reagent at the 3′ terminal.

Each RNA sample underwent hybridization with one gene array to produce the expression data of all genes on the array. We performed eleven independent bacterial cultures and RNA extractions at different times, and collected eleven sets of microarray data for this study. A global normalization scheme is applied so that each array's median value is adjusted to a predefine value (500). The scale factor for achieving this transformed median value for an array is uniformly applied to all the probe set values on a specific array to result in the determined signal value for all the probe sets on the array. In this manner, corresponding probe sets can now be directly compared across arrays.

2.4. Bioinformatic analysis

2.4.1. Gene expression analysis

The gene expression data were analyzed by the program GCOS (GeneChip Operating Software) version 1.4. In the program, the Detection algorithm determines whether a measured transcript is detected (P Call) or not detected (A Call) on a single array according to the detection P-value that is computed by applying the one-sided Wilcoxon's signed rank test to test the discrimination scores (R) against a predefined adjustable threshold τ. The discrimination score calculated for each probe pair is a function of the PM intensity (PMI) and the MM intensity (MMI), as given by

$$\mathrm{R} = \frac{\mathrm{PMI} - \mathrm{MMI}}{\mathrm{PMI} + \mathrm{MMI}}. \tag{1}$$

The parameter τ controls the sensitivity and specificity of the analysis, and was set to a typical value of 0.015, and the detection p-value cutoffs, α_1 and α_2, set to their typical values, 0.04 and 0.06, respectively, according to the Affymetrix system.

2.4.2. Gene prediction

Protein-coding region identification and gene prediction were performed by the programs, GeneMark and GeneMark.hmm [8, 9] (http://exon.gatech.edu/GeneMark), respectively. The prokaryotic version and the *M. tuberculosis* H37Rv genome were selected. Both programs use inhomogeneous Markov chain models for coding DNA and homogeneous Markov chain models for noncoding DNA. GeneMark adopts Bayesian formalism, while GeneMark.hmm uses a hidden Markov model (HMM).

2.4.3. Protein domain search

The Pfam program version 20.0 [11] (http://pfam.wustl.edu) was employed to conduct protein domain search after the input DNA sequence was translated into a protein sequence in six possible frames. The search mode was set to "global and local alignments merged," and the cut-off E-value set to 0.001, which is more stringent than the default value of 1.0. Pfam maintains a comprehensive collection of multiple sequence alignments and hidden Markov models for 8296 common protein families based on the Swissprot 48.9 and SP-TrEMBL 31.9 protein sequence databases.

2.4.4. Homology search

The BLASTx program [12] (http://www.ncbi.nlm.nih.gov/BLAST) was used to identify high-scoring homologous sequences. The program first translated the input DNA sequence into a protein sequence in six possible frames, and then matched it against the nonredundant protein sequence database (nr) in the GenBank and calculated the statistical significance of the matches. The default cut-off E-value was 10.0 but we set it to 1.0×10^{-10}. Potential protein-coding genes are defined based on the bioinformatics criteria consisting of the gene structure, protein coding potential, and presence of ortholog evidence. Orthologs refer to homologs in different strains of *M. tuberculosis*. A typical prokaryotic gene has the following structure: the promoter, transcription initiation, the 5′ untranslated region, translation initiation, the coding region, translation stop, the 3′ untranslated region, transcription stop.

3. RESULTS

We conducted a genome-wide expression analysis on intergenic regions using the Affymetrix GeneChip. Each intergenic sequence is subject to gene prediction and coding potential analysis based on bioinformatics. Each candidate gene is validated by sequence comparison with orthologs among other *Mycobacterium tuberculosis* strains.

To analyze the transcriptional activity of intergenic regions, we collected a set of eleven independent RNA samples from *M. tuberculosis*. Each RNA sample contained the information of genome-wide expression of genes, including those residing in the intergenic regions that have yet to be revealed. The Affymetrix GeneChip was used since it contained encoded intergenic sequences whereas other types of microarray like the cDNA array did not.

3.1. Identification of potential genes in intergenic regions

In our analysis, an intergenic region is assumed to transcribe if there exist transcripts that can bind to the probes encoding that intergenic sequence. The presence or absence of a given transcript is determined in accordance with the detection algorithm of the Affymetrix system. A gene or intergenic region was determined to express (transcriptionally active) only if the derived mRNA was present (P-call) in more than 90% of the collected RNA samples with a detection P-value $<$ .001. The active-transcription status assigned to an intergenic sequence signifies the possible presence of a gene within that sequence. However, if a piece of DNA transcribes into a regulatory RNA instead of mRNA, it should not be considered as a protein-coding sequence. Furthermore, it is not clear how much cross-hybridization can occur between genic and intergenic sequences. To minimize false positives for gene identification, the functional criterion based on expression activity should be strengthened by structural analysis.

Gene structure and coding potential are the two mutually supportive elements in the sequence-based approach to gene prediction. The GeneMark algorithm was applied to an intergenic sequence for checking whether it contained a probable coding region, and the GeneMark.hmm algorithm for predicting a gene within the sequence. The criteria based on the predefined transcriptional evidence, coding potential, and gene prediction yielded 65 candidate genes in the intergenic regions of *M. tb.* H37Rv; their locations in the genome are provided at (http://www.patcar.org/Research/MTB_H37Rv_IG.html).

3.2. Protein domain search

The intergenic sequences that satisfied the criteria based on transcription and predicted gene/coding potential were examined for possessing any domain of known function. Pfam search on the protein sequences of candidate genes showed that twelve of them had a known domain (Tables 1, 2). In these cases, a domain was found within the predicted gene, but there were a few exceptions (i.e., IG398 and IG1140) where a domain was found within the intergenic sequence but outside the predicted gene. The function of a gene may be deducible from its associated domain but cannot be confirmed until there is sufficient evidence from homology or biochemistry.

3.3. Gene function prediction

Identification of orthologs is a reliable means for predicting the function of an unknown gene sequence. BLAST, a bioinformatics program for inferring functional and evolutionary relationships between sequences, was employed to retrieve from sequence databases all proteins that produce statistically significant alignment with a given intergenic sequence under study. The sequences thus obtained are homologous to the query sequence. The highest-scoring homologous sequences with $\geq$ 98% identity consistently turned out to be those belonging to the same strain (H37Rv) or different strains of *Mycobacterium tuberculosis* (e.g., CDC1551, F11, and C) in this analysis.

A homologous sequence found in different strains of the same species often represents an ortholog that shares similar function, whereas a homologous sequence in the same organism could be a paralog that tends to have different function. Paralogs were not found. In fact, given an intergenic sequence, when the BLAST program returned a homologous sequence pertaining to the H37Rv strain, it was actually the same protein-coding sequence contained in the

Table 1: Intergenic sequences in the genome of *Mycobacterium tuberculosis* H37Rv. This list includes intergenic sequences that exhibit gene expression and contain a predicted gene as well as a known domain. The starting and ending positions refer to those in the genome. The strand refers to the coding strand or the strand associated with a higher expression signal. "Exp" is the mean level of the gene expression.

IG	Start	End	Exp	Gene-Start	Gene-End	Strand
IG1061	1485277	1485859	3900	1485311	1485766	−
IG499	731675	731927	2230	731710	731877	+
IG617	882417	882757	1072	882522	882755	+
IG1741	2486986	2487612	698	2486992	2487414	+
IG2500	3571209	3571598	624	3571332	3571586	+
IG2053	2958344	2958905	521	2958346	2958867	+
IG1179	1678903	1679319	502	1678940	1679170	+
IG2522	3600696	3601011	371	3600697	3601009	+
IG1567	2234648	2234988	413	2234650	2234889	−
IG2229	3167800	3168579	237	3168209	3168424	+

Table 2: Each intergenic sequence shown is characterized by its flanking genes or ORFs and the functional domain identified in the translated protein sequence. Most of IGs with a functional domain contain a gene in the reannotated H37Rv genome.

IG	Lt Flank	Rt Flank	Domain	Reannotated H37Rv Gene
IG1061	Rv1322	Rv1323	Glyoxalase	Rv1322A*
IG499	Rv0634c	Rv0635	Ribosomal_L33	Rv0634B
IG617	Rv0787	Rv0788	PurS	Rv0787A
IG398	Rv0500	Rv0501	DUF1713	Rv0500A*
IG1741	Rv2219	Rv2220	RDD	Rv2219A
IG2500	Rv3198c	Rv3199c	Glutaredoxin	Rv3198A
IG2053	Rv2631	Rv2632c	UPF0027	Rv2631*
IG1179	Rv1489c	Rv1490	MM_CoA_mutase	Rv1489A*
IG1140	Rv1438	Rv1439c	TetR_N	None
IG2522	Rv3224	Rv3225c	YbaK	Rv3224B*
IG1567	Rv1991c	Rv1992c	RHH_1	None
IG2229	Rv2856	Rv2857c	cobW	None

* Hypothetical protein.

intergenic sequence, as evident from the fact that they both occupied the same location in the H37Rv genome. This situation arose because the intergenic sequence was taken from the original version of the H37Rv genome while the homologous sequence was based on the later revised version stored in the database. The significance of this finding is twofold. First, a noncoding sequence could be upgraded to one containing a coding region as a result of more research. Secondly, our method based on bioinformatics and transcriptional evidence has correctly predicted these changes in a more time-economical way. The changes refer to IG1061 → (containing) Rv1322A, IG499 → Rv0634B, IG617 → Rv0787A, IG1741 → Rv2219A, IG2500 → Rv3198A, IG2053 → Rv2631, IG1179 → Rv1489A, IG2522 → Rv3224B, IG1291 → Rv1638A, IG398 → Rv0500A, IG2870 → Rv3678A, IG188 → Rv0236A, IG2498 → Rv3196A, IG2591 → Rv3312A, IG595 → Rv0755A, IG1814 → Rv2309A, IG1030 → Rv1290A, and IG2141 → Rv2737A. Here each intergenic region contained an independent gene/CDS with the only exception that part of IG2053 was incorporated in its left-flanking CDS. The presence of a gene structure in an IG and its lack of functional correlation with its adjacent genes suggest that it is not a run-away segment from adjacent genes.

Potential protein-coding genes in our analysis refer to those satisfying the bioinformatics criteria defined earlier. A probable function can be assigned to a candidate gene if it is homologous to another gene of know function, but the strategy of inferring the function of an uncharacterized sequence from its orthologs had limited value in analyzing intergenic data in the present study mainly because most of the found orthologs were hypothetical proteins with unknown function. A candidate gene that contained a known functional domain was not assigned a specific function unless it had an ortholog of known function. Without a specific function assigned, we would term a CDS a hypothetical protein rather than a gene.

The bioinformatics criteria in conjunction with transcriptional evidence revealed potential protein-coding genes with a specific function implied by orthologs in 6 intergenic sequences: IG499, IG617, IG1741, IG2500, IG1567, and IG2229, among which 4 genes had been reported in the *M. tuberculosis* H37Rv genome (Table 2). A hypothetical protein

Table 3: The locations of new hypothetical proteins found in the genome of *Mycobacterium tuberculosis* H37Rv. Each IG listed contains a predicted gene (not shown), whose locations in the genome are given at http://www.patcar.org/Research/MTB_H37Rv_IG.html.

IG	Start	End	Exp	Strand	Orthologs in *M. tuberculosis*
IG914	1271907	1272420	3130	−	MT1178
IG1753	2510255	2510595	1294	−	MT2297
IG2456	3502934	3503389	942	+	MT3222
IG1680	2398405	2398717	912	+	MtubF_01002217, MtubC_01001975
IG2210	3136331	3136616	893	−	MT2896
IG985	1371476	1371774	880	−	MT1266
IG454	665382	665848	782	−	MT0600
IG1989	2869236	2869724	651	−	MT2625, MtubF_01002636, MtubC_01002404
IG3016	4319638	4320700	538	−	MT3957
IG23	31820	32056	520	−	MT0031
IG789	1113582	1114290	505	+	MT1025.2, MtubF_01001043, MtubC_01000775
IG1093	1539210	1539509	502	+	MT1413, MtubF_01001433, MtubC_01001168
IG1670	2387971	2388613	493	+	MtubF_01002203, MtubC_01001961
IG1140	1616348	1616958	492	−	MtubF_01001501, MtubC_01001241
IG1359	1961787	1962225	409	−	MT1777, MtubF_01001795, MtubC_01001544
IG2681	3848802	3849289	407	−	MtubF_01003537, MtubC_01003989
IG717	1016684	1017214	401	−	MT0937
IG1685	2402509	2402974	391	−	MT2201
IG525	767319	767681	384	+	MT0697
IG1652	2364780	2365462	375	−	MT2165
IG1812	2581134	2581761	359	−	MT2367.1
IG1546	2205272	2205579	293	−	MT2013
IG53	68361	68617	266	+	MT0069, MtubF_01000066, MtubC_01003319
IG713	1014123	1014678	254	+	MT0932, MtubF_01000953, MtubC_01000683
IG758	1073272	1073542	249	+	MT0987, MtubF_01001005, MtubC_01000736
IG2313	3317459	3318326	232	+	MT3041.1
IG1087	1530924	1531345	217	−	MT1404, MtubF_01001425, MtubC_01001160
IG54	71558	71818	186	+	MtubF_01000069, MtubC_01003322
IG2849	4092876	4093628	185	+	MT3755
IG2360	3378241	3378707	154	+	MT3103, MtubF_01003110
IG2492	3558343	3559366	151	−	MT3282
IG1498	2141868	2142518	119	−	MT1945
IG2618	3755030	3755947	115	+	MT3456.1
IG331	503123	503493	106	+	MT0431, MtubF_01000431, MtubC_01000146
IG1849	2632074	2632920	102	+	MT2418
IG1560	2225831	2226241	97	−	MT2035
IG2363	3380680	3381371	92	−	MT3106.1
IG841	1178391	1179393	78	+	MT1086, MtubF_01001104, MtubC_01000837

was found in 52 intergenic sequences and 14 among them had been reported in the H37Rv genome. Taken together, there were two genes with a specific function and 38 hypothetical proteins (Table 3) that had not been reported in the H37Rv genome. The two genes mentioned are a DNA-binding protein in the CopG family and a nickle binding GTPase, located in IG1567 and IG2229, respectively (Figure 1). Importantly, 4.3% of intergenic regions exhibiting transcriptional evidence contained a gene in the reannotated H37Rv genome, compared with 1.0% of intergenic regions in the absence transcriptional evidence. The four-fold increase in likelihood in the results suggests that microarray-based transcriptional evidence would facilitate genome-wide gene finding.

4. DISCUSSION

The computational part of the gene prediction problem is dealt with by two classes of algorithms. One is based on sequence similarity while the other based on gene structure and signal is known as *ab initio* prediction. The first class of algorithms, exemplified by BLAST [12], finds sequences (DNA,

protein, or ESTs) in the database that match the given sequence, whereas the second class of algorithm, notably hidden Markov model [8, 9, 13], builds a model of gene structure from empirical data. They both have their own limitations. For instance, the sequence-based approach cannot handle the case of having no homology, and the model-based approach the case of inadequate training data. The method devised in this study would offer a more reliable gene-prediction mechanism by combining sequence alignment, transcriptional evidence, and homology. In particular, the transcriptional activity of a piece of DNA is direct evidence that it is functioning. As the whole H37Rv genome sequence has been intensively searched for genes, transcriptional analysis of intergenic regions could only provide more insight into hidden genes. The integrated method suggested by this study makes sense from our data showing that transcriptional evidence can support finding potential protein-coding genes in the intergenic regions. Thus the idea of combining the evidence from the sequence- and function-based analyses lends itself to not just gene characterization but also gene prediction. Notice, however, genes that are silent in the standard in vitro growth condition are not subject to examination in this study, but the same method can be used generally for gene finding in other genomes and conditions.

We studied the intergenic regions of *M. tuberculosis* H37Rv because of our observation that some of the intergenic regions exhibit expression signals. This observation has little to do with our traditional understanding about promoter and *cis*-regulatory elements since the former is involved in binding of RNA polymerase and the latter in binding transcriptional factors but the DNA-protein binding process dose not require transcription in the intergenic region. Relevant to this discourse is the fact that there are a number of regulatory, noncoding RNAs assuming a distinct role from mRNA, rRNA, and tRNA. Many such RNAs have been identified and characterized both in prokaryotes and eukaryotes and their main function is posttranscriptional regulation of gene expression and RNA-directed DNA methylation [14, 15]. A noncoding RNA has neither a long open reading frame nor a gene structure. The DNA sequence that encodes a noncoding RNA may be viewed as a gene if its regulatory function can be defined. An isolated expression element unaccompanied by a gene structure may hint at noncoding or regulatory RNA. We confirmed that the potential protein-coding genes found in this study did not match any RNA family published in the RNA-families database (www.sanger.ac.uk/Software/Rfam).

New genes continue to be discovered over time, but the accumulated discovery will approach to saturation if the true number of genes is a constant, albeit unknown. Advanced genome annotation technology enables the identification of most, if not all, protein-coding sequences in the genome as soon as it is sequenced. Thus, it is reasonable that the number of new protein-coding sequences due to reannotation is merely 2% of that in the original submission of *M. tuberculosis* genome [16]. Through homology and pattern-based search, most protein-coding sequences with a predicted function have been reported. It is encouraging that we have still been able to find a small number of those in this study. The current knowledge concerning *M. tuberculosis* genes is derived from intensive research in the field involving biological experiments, such as gene deletion and complementation, and bioinformatics analysis. The gap between the existing knowledge about *M. tuberculosis* genes in the genome and our findings in this study can be ascribed to the lack of timely update of genome-annotation with the latest research results in bioinformatics and genomics rather than the inconsistency in stringency of computational parameters used. The integrity and advancement of the knowledge base in genomics would hinge upon the maintenance of complete and accurate information about the whole genome, especially for model organisms, such as *M. tuberculosis* H37Rv.

(1) [Location]: Between Rv1991c and Rv1992c
[Product]: DNA-binding protein, CopG family
[Nucleotide Sequence]: atcgtccatggtttctagcacgcggtatgcgttggccacggcgagggcctccgcttcgtcggtgccatggatgctctctagagccctgtcgatctggcccgtgagcaattgggcgtccagctcgtgcaggtagcgctgcgcagccttcgtgaagaactcggaccgactcatgccgagctcactcgcacgccgcgatacccgatcgaacgtctcatccggcagagaaatagctgtcttcat
[Protein Sequence]: mktaislpdetfdrvsrraselgmsrsefftkaaqrylheldaqlltgqidralesihgtdeaealavanayrvletmdd

(2) [Location]: Between Rv2856 and Rv2857c
[Product]: Nickle binding GTPase involved in regulation of expression urease and hydrogenase
[Nucleotide Sequence]: atggtctcctcggtcaccgagggcaaggacaagccgctgatgtacccggcgacgttccgctcgagggatgtagtgctgctcgacaagatcgacttggtgccctttctggacgccgacgtggacgcgtatatcgcgcatgtccgcgaggtcaacgcagccgcgacgatcctgccgaccagcacgcgcaccggagccggcatggggtcctggtcatga
[Protein Sequence]: mvssvtegkdkplmypatfrsrdvvlldkidlvpfldadvdayiahvrevnaaatilptstrtgagmgsws

Figure 1: New genes with a predicted function found in the genome of *Mycobacterium tuberculosis* H37Rv.

A critical element in this research is the Affymetrix oligonucleotide GeneChip, which allowed us to detect the gene expression of the intergenic regions in *M. tuberculosis* H37Rv. The Affymetrix system can compute the absolute signal intensity of mRNA hybridized on the array in a single condition as well as the signal ratio between two conditions. The built-in statistical algorithm arrives at the so-called detection *P*-value that determines the presence or absence of any given mRNA. In contrast, the cDNA microarray, another major platform, generally does not indicate whether and to what extent a gene expresses in each condition. While there exist a couple of other types of oligonucleotide microarray, only the Affymetrix array implements the probes for interrogating intergenic sequences in the H37Rv genome. As an additional strength, the Affymetrix array is designed to minimize cross-hybridization by using unique oligonucleotide probes and the pair of PM (perfect-match) and MM (mismatch) probes. The cross-hybridization of related or overlapping gene sequences often contributes to false positive signals, especially in the case when long cDNA sequences are used as probes. A study demonstrated that the Affymetrix GeneChip produced more reliable results in detecting changes in gene expression than cDNA microarrays

[17]. Thus, the choice of the Affymetrix GeneChip for this study is well justified. To validate genome-wide microarray data, a basic means is to demonstrate a high correlation between the data of duplicate experiments [18]. In the present study, the correlation between any pair of the gene expression data derived from independent RNA samples is > .9. In addition, PCR analysis has been performed to verify that the Affymetrix Genechip system worked properly in our prior work [19, 20].

5. CONCLUSION

Current computational programs for gene prediction have no guarantee to identify all genes in a sequenced genome because the knowledge about gene structure has yet to be perfected. Genome reannotation using the same kind of heuristics offers limited help unless its predictive power has been improved. Reannotation based on new experimental evidence that trickles in at its own pace is probably slow.

We conducted a genome-wide analysis using the Affymetrix GeneChip to explore genes contained in the intergenic sequences of the *M. tuberculosis* H37Rv genome. Potential protein-coding genes were determined according to the bioinformatics criteria constituted by the gene structure, protein coding potential, and the presence of ortholog evidence. The bioinformatics criteria in conjunction with transcriptional evidence have led to the discovery of genes with a specific function, such as a DNA-binding protein in the CopG family and a nickle binding GTPase, as well as hypothetical proteins that have not been reported in the *M. tuberculosis* H37Rv genome. This work has demonstrated that microarray-based transcriptional evidence would help gene finding on the genomic scale.

ACKNOWLEDGMENTS

This work is supported by National Institutes of Health under the Grant HL-080311 and the Centers of Disease Control and Prevention. The authors would like to thank CDC for the use of the facilities and UCI for providing service for microarray hybridization. They also thank Thomas R. Gingeras at Affymetrix, Inc. for designing *Mycobacterium tuberculosis* GeneChip. Bacterial culture and RNA isolation were performed by Pramod Aryal.

REFERENCES

[1] S. T. Cole, R. Brosch, J. Parkhill, et al., "Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence," *Nature*, vol. 393, no. 6685, pp. 537–544, 1998.

[2] R. Overbeek, T. Begley, R. M. Butler, et al., "The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes," *Nucleic Acids Research*, vol. 33, no. 17, pp. 5691–5702, 2005.

[3] G. H. Van Domselaar, P. Stothard, S. Shrivastava, et al., "BASys: a web server for automated bacterial genome annotation," *Nucleic Acids Research*, vol. 33, Web Server issue, pp. W455–W459, 2005.

[4] P. Stothard and D. S. Wishart, "Automated bacterial genome analysis and annotation," *Current Opinion in Microbiology*, vol. 9, no. 5, pp. 505–510, 2006.

[5] P. Nielsen and A. Krogh, "Large-scale prokaryotic gene prediction and comparison to genome annotation," *Bioinformatics*, vol. 21, no. 24, pp. 4322–4329, 2005.

[6] J.-M. Lee, S. Zhang, S. Saha, S. Santa Anna, C. Jiang, and J. Perkins, "RNA expression analysis using an antisense *Bacillus subtilis* genome array," *Journal of Bacteriology*, vol. 183, no. 24, pp. 7371–7380, 2001.

[7] D. Zheng, Z. Zhang, P. M. Harrison, J. Karro, N. Carriero, and M. Gerstein, "Integrated pseudogene annotation for human chromosome 22: evidence for transcription," *Journal of Molecular Biology*, vol. 349, no. 1, pp. 27–45, 2005.

[8] A. V. Lukashin and M. Borodovsky, "GeneMark.hmm: new solutions for gene finding," *Nucleic Acids Research*, vol. 26, no. 4, pp. 1107–1115, 1998.

[9] J. Besemer and M. Borodovsky, "GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses," *Nucleic Acids Research*, vol. 33, Web Server issue, pp. W451–W454, 2005.

[10] M. A. Fisher, B. B. Plikaytis, and T. M. Shinnick, "Microarray analysis of the *Mycobacterium tuberculosis* transcriptional response to the acidic conditions found in phagosomes," *Journal of Bacteriology*, vol. 184, no. 14, pp. 4025–4032, 2002.

[11] R. D. Finn, J. Mistry, B. Schuster-Böckler, et al., "Pfam: clans, web tools and services," *Nucleic Acids Research*, vol. 34, Database issue, pp. D247–D251, 2006.

[12] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[13] C. Burge and S. Karlin, "Prediction of complete gene structures in human genomic DNA," *Journal of Molecular Biology*, vol. 268, no. 1, pp. 78–94, 1997.

[14] V. A. Erdmann, M. Z. Barciszewska, A. Hochberg, N. de Groot, and J. Barciszewski, "Regulatory RNAs," *Cellular and Molecular Life Sciences*, vol. 58, no. 7, pp. 960–977, 2001.

[15] A. S. Pickford and C. Cogoni, "RNA-mediated gene silencing," *Cellular and Molecular Life Sciences*, vol. 60, no. 5, pp. 871–882, 2003.

[16] J.-C. Camus, M. J. Pryor, C. Médigue, and S. T. Cole, "Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv," *Microbiology*, vol. 148, no. 10, pp. 2967–2973, 2002.

[17] J. Li, M. Pankratz, and J. A. Johnson, "Differential gene expression patterns revealed by oligonucleotide versus long cDNA arrays," *Toxicological Sciences*, vol. 69, no. 2, pp. 383–390, 2002.

[18] J. L. DeRisi, V. R. Iyer, and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, no. 5338, pp. 680–686, 1997.

[19] L. M. Fu, "Exploring drug action on *Mycobacterium tuberculosis* using affymetrix oligonucleotide genechips," *Tuberculosis*, vol. 86, no. 2, pp. 134–143, 2006.

[20] L. M. Fu and T. M. Shinnick, "Genome-wide exploration of the drug action of capreomycin on *Mycobacterium tuberculosis* using Affymetrix oligonucleotide GeneChips," *Journal of Infection*, vol. 54, no. 3, pp. 277–284, 2007.