Published online 15 November 2007

Nucleic Acids Research, 2007, Vol. 35, No. 21 ***e148***
doi:10.1093/nar/gkm918

A bioinformatic filter for improved base-call accuracy and polymorphism detection using the Affymetrix GeneChip® whole-genome resequencing platform

Gagan A. Pandya[1], Michael H. Holmes[1], Sirisha Sunkara[1], Andrew Sparks[2], Yun Bai[1], Kathleen Verratti[1], Kelly Saeed[1], Pratap Venepally[1], Behnam Jarrahi[1], Robert D. Fleischmann[1] and Scott N. Peterson[1,*]

[1]Pathogen Functional Genomics Resource Center, The Institute for Genomic Research at the J. Craig Venter Institute, Rockville, MD 20850, USA and [2]Affymetrix, Inc., Santa Clara, CA 95051, USA

Received August 13, 2007; Revised September 21, 2007; Accepted October 9, 2007

ABSTRACT

DNA resequencing arrays enable rapid acquisition of high-quality sequence data. This technology represents a promising platform for rapid high-resolution genotyping of microorganisms. Traditional array-based resequencing methods have relied on the use of specific PCR-amplified fragments from the query samples as hybridization targets. While this specificity in the target DNA population reduces the potential for artifacts caused by cross-hybridization, the subsampling of the query genome limits the sequence coverage that can be obtained and therefore reduces the technique's resolution as a genotyping method. We have developed and validated an Affymetrix Inc. GeneChip® array-based, whole-genome resequencing platform for *Francisella tularensis*, the causative agent of tularemia. A set of bioinformatic filters that targeted systematic base-calling errors caused by cross-hybridization between the whole-genome sample and the array probes and by deletions in the sample DNA relative to the chip reference sequence were developed. Our approach eliminated 91% of the false-positive single-nucleotide polymorphism calls identified in the SCHU S4 query sample, at the cost of 10.7% of the true positives, yielding a total base-calling accuracy of 99.992%.

INTRODUCTION

Detection, identification and typing of infectious microorganisms are crucial in many areas of basic and translational research. There is an urgent need for accurate, high-resolution microbial genotyping methods, in spite of the existence of a number of widely used typing methods. That is because the current methods offer only low-resolution and limited genotyping. Single nucleotide polymorphisms (SNPs), horizontal gene transfer and/or intragenic recombination are all known to lead to variations and genome plasticity in an organism. The diversity of an organism can be explicitly understood provided these events are identified and analyzed quantitatively and qualitatively at the whole-genome level.

Whole-genome sequencing is probably the most accurate and reliable method to identify and type strains of a species. SNPs and other polymorphisms that serve as informative genetic characters are globally determined and therefore enable a more complete evaluation of inferred evolutionary relationships. The cumulative differences between two or more sequences provide a larger framework upon which reliable phylogenies may be established. Strain genotypes that are built upon SNP variation are highly amenable to evolutionary reconstruction and can be readily analyzed in a phylogenetic and a population genetic context to: (i) assign unknown strains into well-characterized clusters, (ii) reveal closely related siblings of a particular strain and (iii) examine the prevalence of a specific allele in a population of closely related strains that may in turn correlate with phenotypic features of the infectious agent (1). More directly and for a variety of purposes such as forensic investigations or epidemiological investigations, SNPs provide potential markers for the purpose of strain identification.

The increased availability of complete DNA sequence data for a large number of reference genomes has elevated the value of resequencing methods. Whole-genome DNA resequencing data offer several advantages over existing DNA-based typing methods. We have exploited the

*To whom correspondence should be addressed. Tel: +1 301 795 7539; Fax: +1 301 838 0208; Email: scottp@jcvi.org
Present address:
Andrew Sparks, Complete Genomics, 658 North Pastoria Avenue, Sunnyvale, CA 94085, USA

modern high-density oligonucleotide arrays as an alternative to the classical ABI sequencing approach towards achieving whole-genome sequence information. Mockler *et al*. (2) recently provided an excellent overview of applications of DNA tiling arrays for whole-genome analysis that includes genome resequencing, genotyping and polymorphism discovery. The whole-genome, array-based resequencing and SNP identification approach is simple and time efficient, enabling high-resolution analysis of a number of strains in a matter of days.

Several studies have been published describing the principles of resequencing array technology, emphasizing its application for genotyping in both prokaryotes and eukaryotes (3–10). The underlying Affymetrix® base-calling software (4) used in the reported studies, although powerful, is limited in its capacity to account for some genome-scale induced artifacts. For example, deletions in the sample DNA relative to the reference sequence can cause poor hybridization performance, resulting in a mixture of no-calls and false-positive SNP calls in the affected regions (the low homology effect); the large population of target DNA fragments in a whole-genome sample may contain sequences capable of high-efficiency hybridization with more than one of the probe pairs, resulting in a false-positive SNP call (the alternate homology effect) and the local destabilizing effect of genuine SNPs in the sample leads to false-positive calls at adjacent genome locations (the footprint effect).

Here, we discuss some of these systematic effects and report a novel bioinformatic filtering approach to mitigate them. Our approach increased base-call accuracy and significantly reduced false positives for whole-genome resequencing of *Francisella tularensis* using Affymetrix, Inc. GeneChip® 300 K resequencing arrays. *Francisella tularensis* is the causative agent of tularemia in humans and a select A agent. Our approach will lead to a more reliable global SNP identification and genotyping platform. Genotyping of strains globally at a single nucleotide resolution will prove useful for both basic and applied areas of biodefense and infectious disease research.

MATERIALS AND METHODS

Francisella tularensis genomic DNA

Genomic DNAs of *F. tularensis* reference strains LVS and SCHU S4 were obtained from Dr Luther Lindler at Walter Reed Army Research Institute, MD. Dr C. Ben Beard at The Centers for Disease Control and Prevention (CDC), Fort Collins, CO and Dr Mark J. Wolcott of U.S. Army Medical Research Institute of Infectious Diseases (USAMRIID) Frederick, MD, provided us with genomic DNAs of other *F. tularensis* strains (Supplementary Table 1), used in batch analysis of data. Genomic DNA samples were stored at −80°C. The whole-genome resequencing was performed in duplicate for all sequences used in the batch analysis and query strains were sequenced in quadruplicate.

Francisella tularensis custom resequencing array set

The basis of the Affymetrix GeneChip® resequencing by hybridization is shown in Supplementary Figure 1. Each queryable base position in the reference sequence (non-repetitive sequence) was represented by eight 25 nucleotide probes or 'features' that define a locus. For each locus, four probes were designed to query the forward strand and four probes represented the reverse complementary strand. The forward and reverse probes were identical at all base positions except the central (13th) position, where the reference base and each of the three alternative bases were represented. The next locus represented probes that were shifted by one base and placed a new base in the central query position. The tiling of loci along a genome sequence effectively allowed for base calls to be derived for each base represented on the array. High confidence base calls were made by virtue of two features within each locus hybridizing with greater efficiency than alternative features (e.g. an A call on the forward strand was accompanied by a T call on the reverse strand). The Affymetrix GSEQ software assigned a quality score to each base call that combined hybridization information from both the forward and reverse strands.

The *F. tularensis* GeneChip® set was designed on the basis of the DNA sequence of strains LVS (GenBank Accession: AM 233362) and SCHU S4 (GenBank Accession: AJ 749949) available at http://cmr.tigr.org. Sequences of plasmids, pOM1 (GenBank Accession: NC 002109) and pFNL10 (GenBank Accession: NC 004952), were obtained from the NCBI database (http://www.ncbi.nlm.nih.gov/). The LVS sequence used in this study was obtained from The Microbial Genomics group, Lawrence Livermore National Laboratory, Los Alamos, prior to its submission to NCBI. This sequence differs from the submitted sequence by 13 insertions and deletions (indels, 12 single base and 1 two base), and 12 variant base calls. All but four of these differences lie in repeat regions that were excluded from our design. The remaining differences are a single base insertion in the final sequence near the start of one of the fragments (or instructions) on the array, and three single base call changes. A merged sequence was constructed based on these genomic and plasmid sequences for the purposes of GeneChip® design. The *F. tularensis* LVS and SCHU S4 genomes are 1 895 998 and 1 892 819 bp, respectively. An *in silico* analysis was performed to identify unique sequences from SCHU S4 (ranging from 1 bp to 11 086 bp) that were appended to the LVS sequence along with plasmid pOM1 sequence and unique regions from pFNL10. There are 12 869 bp of sequences unique to LVS relative to SCHU S4 and 42 369 bp present in SCHU S4 but not LVS. In total, this analysis defined 1 943 751 bp of *F. tularensis* sequence. We used the MUMmer tool set (http://mummer.sourceforge.net/) and repeatFinder [based on REPuter(c) Copyright University of Bielefeld, Germany (http://www.genomes.de/)] to identify 170 356 and 139 560 bp of repetitive sequence in LVS and SCHU S4, respectively. A total of 179 193 bp. (9.22%) of repetitive sequence were excluded from the design, resulting in 1 764 558 queryable bases (91% of the

F. tularensis genome) for resequencing by hybridization. A total of 1 769 695 bp were submitted for chip production by adding back 5137 bp from the immediate flanks of excluded repeats as padded bases. This sequence was tiled onto a set of six CustomSeq 300 K GeneChips® by Affymetrix, Inc. (Santa Clara, CA), consisting of 14 125 688 individual probes. A maximum of 303 366 bases of double-stranded DNA can be resequenced on a 300 K array.

Whole-genome amplification

Francisella tularensis genomic DNA was subjected to whole-genome amplification (WGA) by multiple displacement amplification (MDA). MDA was performed using ϕ29 DNA polymerase using the Repli-g kit (Qiagen Inc, Valencia, CA) in 50 µl reaction volumes as follows. Genomic DNA (10 ng) in 2.5 µl of 1× TE was denatured at room temperature for 3 min by the addition of an equal volume of 50 mM KOH, 1.25 mM EDTA (pH 8.0). The solution was neutralized by the addition of 5 µl of 1 M *Tris*–HCl, pH 4.0. A master mix (40 µl) containing 50 µM exonuclease-resistant random hexamers, 1 mM of each dNTPs, ϕ29 DNA polymerase (800 U/ml final concentration) and yeast pyrophosphatase (1 U/ml final concentration) was added. The reaction was incubated at 30°C for 16 h in a thermocycler PTC-225 (MJ Research, Waltham, MA). The reactions were terminated by heating at 65°C for 3 min. The amplified DNAs were purified using 96-well purification microplates (Millipore, Billerica, MA). The DNAs were eluted for 1 h at room temperature on a shaker in 100 µl of 1× TE. The amplified DNAs were examined on agarose gels and DNA concentrations were determined using Pico Green dsDNA quantitation kit (Invitrogen—Molecular Probes, Carlsbad, CA) using calf thymus DNA (Sigma-Aldrich, St Louis, MO) as a standard. The sample fluorescence was measured using a fluorescence microplate reader (TECAN, San Jose, CA) at excitation of 480 nm and measuring emission at 520 nm. Amplified DNA yields were typically 30–40 µg. The 7.5 kb plasmid DNA used as a positive control for resequencing (Tag IQ-EX template) was PCR amplified using the primers and conditions suggested in the CustomSeq® kit (Affymetrix, Inc., Santa Clara, CA).

DNA fragmentation, labeling and hybridization

GeneChip® resequencing assay kit (Affymetrix, Inc., Santa Clara, CA) was used for DNA fragmentation and labeling of amplified *F. tularensis* DNA. Briefly, 12 µg of amplified DNA was fragmented in a 300 µl reaction and 12 µl of fragmentation reagent (0.15 U/µl, Affymetrix, Inc., Santa Clara, CA) for 20 min at 37°C. Reactions were terminated by heat treatment at 95°C for 15 min. Labeling reactions, prehybridization, hybridization, washing, staining and scanning of the arrays were performed as described (https://www.affymetrix.com/support/downloads/manuals/customseq_protocol.pdf) by Affymetrix, Inc. Chips were washed and stained on the GeneChip® fluidics station 450 using the pre-programmed Mapping 100Kv1_450 wash protocol and scanned with GeneChip® Scanner 3000 (Affymetrix, Inc., Santa Clara, CA).

Raw data acquisition

The Affymetrix GeneChip® Sequence Analysis Software (GSEQ) Version 4.0 was used to analyze hybridization results and to obtain raw data. The GSEQ software implements a batch analysis approach that allows correction for background signals by comparing signal intensities at each base position across a set of samples that makes up a batch. We used a batch size of 16. Fifteen samples were selected to maximize sample diversity and establish a background set. Each query sample was added individually to the batch for analysis. Sample diversity was determined by phylogenetic analysis (Supplementary Figure 2) using the resequencing results from 40 samples analyzed in non-batch mode. Except when noted, we used the following analysis parameters in the GSEQ software: default setting for filter conditions, haploid genome model, trace threshold and sequence profile threshold in final reliability rules was used. The quality score threshold was set to 12.000 in the base-calling parameters and the call rate cutoff across samples was turned off (0.0000). The GSEQ software produces a CHP file containing the base call, the corresponding reference base and a quality score for each locus on the resequencing chip.

Bioinformatic filters

The GSEQ software produces a CHP file containing the base call, the corresponding reference base and a quality score for each locus on the resequencing chip. Our bioinformatic filters consist of a set of Perl scripts that operate on the CHP files and produce a list of high-confidence SNP calls from the larger raw set of SNP calls present in those files. The scripts are available for download from our website (http://pfgrc.jcvi.org/presentations/data/snp_filter_scripts/snp_filter_scripts.shtml). Each filter serves to reduce the number of candidate SNPs. The output of one filtering step becomes the input for the next.

The availability of a published complete genome sequence for the SCHU S4 strain allowed us to predict the set of expected SNP calls that should occur as the result of hybridizing the SCHU S4 sample to the LVS portion of the chip set. Consequently, we were able to characterize the SNP calls resulting from these experiments as either true positives or false positives. We used this information to parameterize and validate our filter algorithms.

The first filter applied, referred to as the low-homology filter (mask_low_homology.pl), seeks to identify regions that performed poorly as a result of deletions in the sample relative to the reference sequence. It scans the base calls from the CHP files to identify regions of adjacent positions that are rich in no-calls and SNP calls. It uses a sliding-window approach, first looking at windows of 50-base length (user specified) for regions whose content of no-calls plus SNP calls comprises 60% or greater of the specified window size. Upon encountering such regions, the algorithm uses a 10-base window to examine the

Figure 1. Representation of the 'alternate homology effect'. Query location is shown in bold and mismatches are shown in red. Chip oligonucleotides and sample DNA alignment at SNP location is shown. The top pair represents a sample DNA sequence perfectly matching a reference probe. The next pair illustrates a sample DNA sequence partially matching a SNP probes and therefore capable of hybridizing with high efficiency to the SNP probe pair.

sequence at higher resolution, so that the proximal breakpoint of the low homology region (generally a deletion) is properly defined. The extent of the region is determined by expanding the region using a 50-base length window as far as possible. Once the region limits are determined, the algorithm uses a 10-base window to map the breakpoint of the distal end of the deletion. SNP calls that occur within the defined low-homology region are removed from the list of high-confidence SNP calls. (The coordinates of the low-homology regions themselves are also interesting, as they represent areas of the reference sequence that are not represented in the sample.) The window sizes, and the required percentage of no-calls plus SNP calls, are parameterized and controlled by command-line options.

The next script, referred to as the alternate homology filter, is important particularly when resequencing DNAs of higher complexity. The query DNA sample may contain sequence capable of hybridizing with high efficiency to more than one probe pair at a locus on the array. The occurrence of such sequence, referred to as the 'alternate homology effect', is illustrated in Figure 1. The ability of the GSEQ software to make a base call at any particular locus is dependent upon the relative signal strength of the best forward and reverse probe being above a certain threshold compared with the next best signal at that locus. When a locus contains two strongly hybridizing probe pairs, the GSEQ software may make a SNP call, a reference base call or a no-call ('N'), depending on the relative signal strengths of the probe pairs. In practice, we have found that the alternate sequence need not match the probe sequence over the entire 25-base length in order to cause a spurious base call. A whole-genome DNA sample will naturally contain a larger population of distinct 25-mer sequences than a sample composed of a small subset of the genome. For this reason, the alternate homology effect is significant in the context of whole-genome hybridization. This problem led us to develop an algorithm to identify and filter out base miscalls that resulted from this phenomenon.

In general, the DNA sequence of query strains is not known. Therefore, it is not possible to identify locations that are subject to alternate homology effects with total confidence. However, an underlying assumption of the resequencing method is that the query DNA is similar

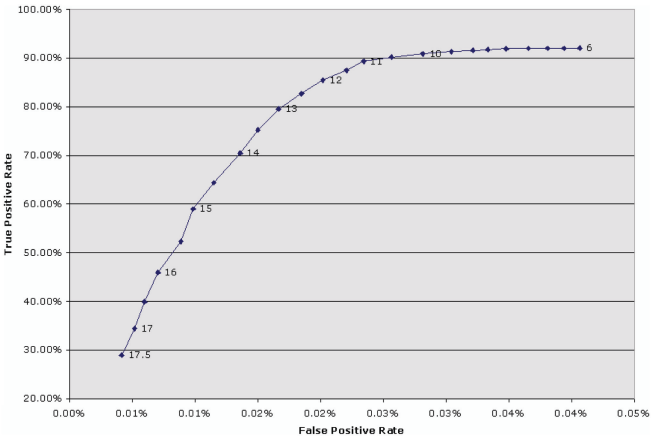

Figure 2. ROC curve showing the effect of different delta binding energy threshold values on the true positive and false positive rates. The values on the line graph are the delta energy values.

to the reference sequence represented on the chip. Our approach to the problem of the alternate homology effect exploits this assumption. For each SNP observed in the raw results, we search for any alternate sequences within the reference sequence that could account for the SNP call. The difference in binding energy between the alternate (SNP) sequence and the reference sequence is used to differentiate between likely artifacts and genuine SNPs. (Calculated binding energy is a much more sensitive predictor of actual binding potential than Tm.)

The alternate homology filter identifies SNP calls that may have arisen as a result of this effect. For each SNP call in the analyzed results, the SNP 25-mer probe sequence is used to search for all perfect, ungapped alignments of at least 13 bases with the LVS genome sequence. The requirement of a minimum alignment length of 13 bases guarantees that the SNP base will be included in all alignments found. The program ExamineSNPs.pl examines the SNP alignments and calculates the binding energies, using the MUMmer package to obtain the sequence alignments (11) and the binding energy calculator from the ArrayOligoSelector package (12) for the binding energy calculations. The alignment representing the highest binding energy is selected and compared with the free energy of binding of the reference 25-mer to its reverse complement. If the difference between these two binding energies is $\leq$11.5 kcal/mol, the SNP call is assumed to be an artifact of the alternate sequence homology and it is removed from the list of high-confidence SNP calls. The set of SNP calls from the hybridization of a SCHU S4 sample was used to determine the threshold binding energy difference that identifies probable alternate homology artifacts. A delta binding energy threshold of 11.5 kcal/mol was chosen based on the effect of different threshold values on the false-negative and false-positive calls (Figure 2).

The next filter in our pipeline is a quality filter that simply eliminates SNP calls that have been assigned low quality scores by the GSEQ software. The quality score is based on the difference in signal intensity between the highest intensity probe pair and the next highest intensity

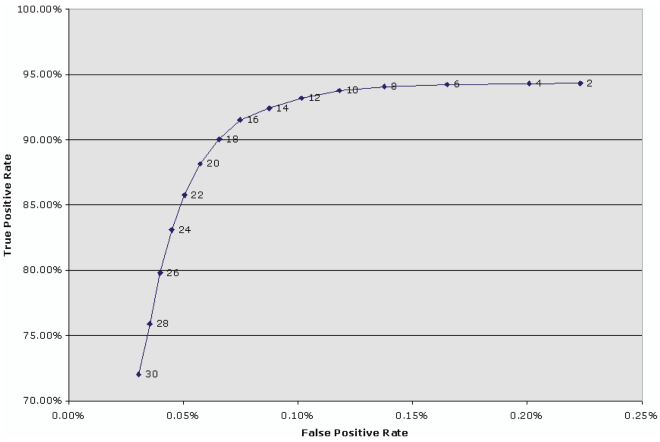

Figure 3. ROC curve illustrating the effect of different quality threshold values on the true positive and false positive rates. The GSEQ quality score threshold was set to 3.0, and our quality filter was applied using different threshold values shown on the line graph.

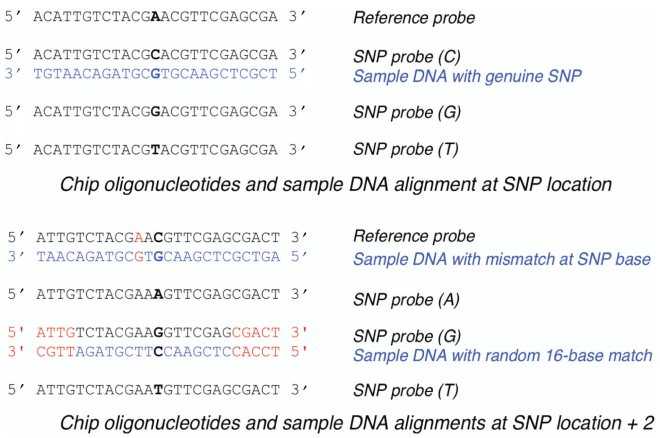

Figure 4. Representation of the 'footprint effect'. Query locations are in bold and mismatches are shown in red. Chip oligonucleotides and sample DNA alignments at SNP location (central 13th position) and SNP location plus two bases are shown.

pair at a particular locus (4), so calls with low quality scores are more likely to be incorrect than high-scoring calls. We have found that filtering out SNP calls with quality scores less than 12.0 removes a large number of false positives, at a relatively small cost in terms of true positives rejected. A receiver operating characteristic (ROC) curve that illustrates the effect of different quality threshold values is shown in Figure 3. (For the analysis in Figure 3 only, we used our own quality filter in preference to the quality filter in the GSEQ software, so that we could easily test the effect of different quality thresholds. For all other analyses, the quality filter incorporated in GSEQ was used. The GSEQ software is run before our filters, so the quality filter was actually the first filter applied, except in the case of Figure 3.)

The remaining SNP calls are next put through the footprint effect filter. The occurrence of a real SNP in a query sample results in a destabilizing effect on 25-mers in the immediate vicinity of the SNP. This artifact, called the 'footprint effect', is illustrated in Figure 4. The locus on the resequencing chip at which the SNP occurs contains two probes that hybridize perfectly with the sample over the entire 25-base length of the forward- and reverse-complement probes (only the forward strand is shown in the figure). However, at adjacent loci on the chip, which represent base positions near the SNP base, there are no probes that hybridize perfectly with the sample DNA. This is because, in general, the chip design tiles probes based on a single reference sequence, which does not contain the SNP base. As a result, the probes on the chip that represent reference sequence positions within 12 bases of the SNP location will all contain at least a single-base mismatch with the sample DNA. This mismatch decreases the reference probe hybridization intensities and increases the likelihood that an alternate sequence from a second location in the sample DNA will hybridize more strongly to a non-reference probe pair. This results in a mixture of reference calls, SNP calls and no-calls at the loci within 12 base positions adjacent to a genuine SNP, with reference calls predominant. This effect is exacerbated when two genuine SNPs occur within the same 25-base window. The footprint effect, like the alternate homology effect, is expected to be more pronounced in the context of a whole-genome hybridization, because of the larger number of hybridization targets in the sample.

The footprint effect filter algorithm assumes that a genuine SNP is most likely to cause spurious SNP calls at locations within 10 bases on either side of the genuine SNP. Any SNP call that occurs more than 10 base positions from the nearest neighboring SNP call is assumed to be valid, and any SNP call that has one or more neighbors within 10 base positions is subjected to the filter. Since any number of consecutive SNP calls within 10 base positions of each other may occur in the data, this filter is implemented as a recursive algorithm. For each list of consecutive SNP calls that each lies within 10 bases of its neighbors, the algorithm identifies the SNP call having the highest quality score. That SNP call is accepted as valid, and its immediate neighbors are removed from the list of high-confidence SNP calls. This action may break the original list of neighboring SNP calls into two separate lists. All resulting lists are processed recursively in the same way, until all of the SNP calls have been accepted or rejected. This algorithm is implemented in the RemoveFootprintEffect.pl Perl program.

Finally, a filter referred to as the replicate combination filter is applied, wherein results from two independent replicates are combined and SNPs present in both the experiments are accepted. This step is performed after the other filters have been applied to individual data sets generated for any single sample.

SNP validation for *F. tularensis* reference strains

The candidate SNPs or unexpected base calls (290) were validated for the reference LVS strain by ABI sequencing. Unexpected SNP calls as well as a subset of false-negative calls (562 total) for the SCHU S4 strain were also validated. Primer pairs flanking miscalled or SNP bases were designed containing M13 universal forward- or reverse

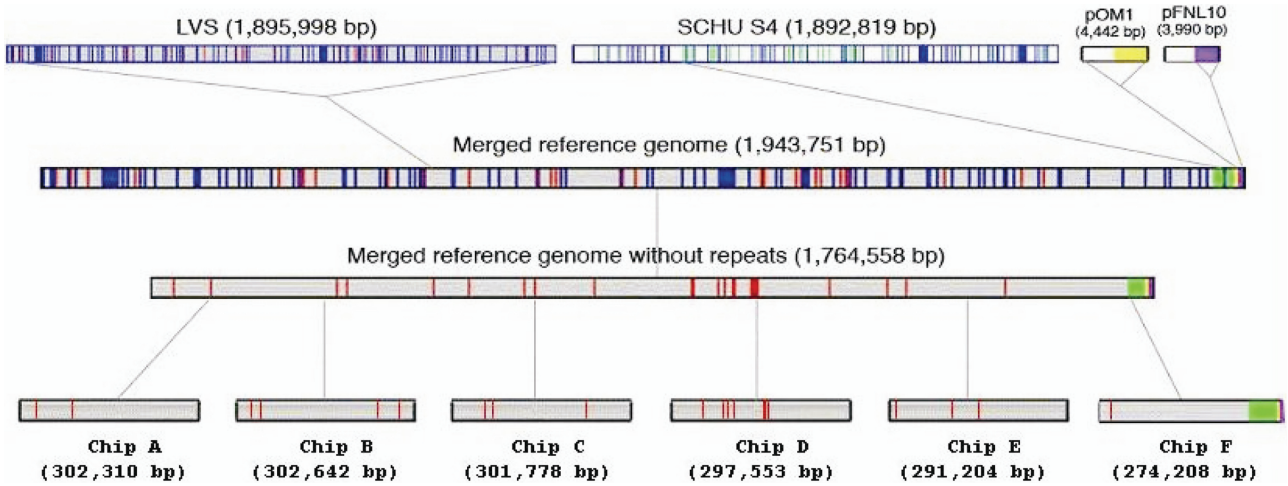

Figure 5. Schematic representation of whole genome resequencing array set design. Blue vertical lines indicate repeats in the genomes. Unique sequences for LVS and SCHU S4 are shown as red and green vertical lines, respectively. Similarly, yellow and purple vertical lines represent unique sequences from plasmids pOM1 and pFNL10, respectively.

sequence at their 5′ end using Primer3 software (13). The Primer3 parameters used for the primer design were as follows: PRIMER_MIN_SIZE = 18
PRIMER_MAX_SIZE = 30
PRIMER_MIN_TM = 60
PRIMER_MAX_TM = 65
PRIMER_PRODUCT_SIZE_RANGE = 150–400
PRIMER_OPT_SIZE = 25
TARGET = 250 150
PRIM_NUM_RETURN = 1

PCR amplification was performed in 50 µl reaction volume using the Takara LA PCR kit, V2.1 (Takara Mirus Bio, Madison, WI) as per the manufacturer's instructions with the following cycling conditions: 95°C for 1 min, followed by 95°C for 30 sec, 56°C for 30 sec and 68°C for 30 sec; for 29 cycles and finally 72°C for 2 min. Amplicons were purified through 96-well purification microplate (Millipore, Billerica, MA) as per the manufacturer's instructions. Purified PCR products were sequenced at the Joint Technology Center (JTC), Rockville, MD, using M13 universal primers.

RESULTS

A schematic representation of the chip design used to represent the *F. tularensis* genome is depicted in Figure 5. The *F. tularensis* LVS genome sequence defined our reference and was represented on chips A–E and the majority of chip F. Unique sequences present in strain SCHU S4, together with two plasmid sequences, were added to the remainder of chip F. Our chip design, based on sequence information from two strains, enables coverage of a large number of strains. Approximately 91% of the *F. tularensis* double-stranded unique genome can be resequenced with this design from strains belonging to *holarctica* (type B) and *tularensis* (type A) subtypes.

Most array-based resequencing studies use purified PCR products rather than whole-genome DNA as described here. We wished to test whether, despite their additional complexity, genomic DNA samples could be reliably resequenced and what specific impediments might be encountered. Hybridization of a whole-genome sample on an Affymetrix® resequencing array platform can lead to incorrect base calls due to a number of systematic effects. Most of these adverse effects were predictable and therefore could be accounted for through application of specific bioinformatic data processing. These effects, and the bioinformatic filters, are described in Materials and Methods section.

Raw reference versus query sample data

In order to determine the base-calling frequency and accuracy of the Affymetrix platform, we performed a series of hybridizations using the LVS reference genomic DNA represented on the resequencing chips and the standard Affymetrix data processing methods. These data are summarized in Table 1 and excludes data from the SCHU S4 and plasmid-specific portions of the resequencing chips. The resequencing results from two experiments on the LVS query sample yielded a call rate $\geq$98.178% and 167 and 177 base calls, respectively, that differed from the chip reference sequence. ABI sequencing of PCR amplicons confirmed that only three of these candidate SNPs were actual differences between the reference genome sequence and our LVS genomic DNA. These three expected SNP locations correspond to sequencing errors in the LVS reference sequence that have since been corrected in the published sequence for LVS (GenBank Accession: AM 233362). These results represent the approximate upper limit of the technology's performance in our hands. It was anticipated that less than optimal results would occur when utilizing the GeneChips® to sequence a non-identical query genome. To assess this assumption, we performed hybridizations using another strain (SCHU S4) for which complete DNA sequence information was available.

The raw resequencing data ($N = 2$) for both the LVS and the SCHU S4 strains, using Affymetrix data

Table 1. Raw resequencing results for *F. tularensis* LVS query against *F. tularensis* LVS reference

Expt. No.	Array	Bases/array	Bases called	Call cate (%)	SNPs	% SNPs of called bases	True–positive SNPs (expected/detected)
007	A	301 470	300 490	99.675	49	0.016	0/0
	B	302 018	297 231	98.415	23	0.008	0/0
	C	301 394	296 530	98.386	27	0.009	0/0
	D	296 905	291 127	98.054	33	0.011	0/0
	E	290 100	282 102	97.243	18	0.006	1/1
	F	234 779	227 722	96.994	17	0.007	2/2
	Total	1 726 666	1 695 202	98.178	167	0.010	3/3
013	A	301 470	300 267	99.601	54	0.018	0/0
	B	302 018	296 718	98.245	25	0.008	0/0
	C	301 394	296 372	98.334	29	0.010	0/0
	D	296 905	290 472	97.833	32	0.011	0/0
	E	290 100	284 220	97.973	20	0.007	1/1
	F	234 779	229 583	97.787	17	0.007	2/2
	Total	1 726 666	1 697 632	98.318	177	0.010	3/3

The results shown are for the LVS sample, using the Affymetrix-recommended batch analysis parameters, including a quality score threshold of 12 and a call rate cutoff of 0.5. Those portions of chip F that represent the SCHU S4 reference and the plasmids were excluded from this analysis. Therefore, this represents the performance of the system under ideal circumstances: the chips are challenged with a sample that is essentially identical to the chip reference.

processing methods, are shown in Table 2. These results used a call rate cutoff value of zero. The recommended value of 0.5 is appropriate in cases where the samples in the batch are highly homogeneous. In our case, the batch was chosen for maximum diversity, and an arbitrary requirement of some minimum fraction of calls across the batch would have resulted in unnecessary loss of data at many locations in the query sample. The results in Table 2 illustrate that the performance of the platform was dependent on the similarity between the query genome and the reference content of the resequencing chip. The number of miscalled bases for SCHU S4 was larger than that observed for LVS; however, the raw data still apparently had a good overall call rate and base-calling efficiency.

Performance of bioinformatic filters

The effects of each filtering step on the base-calling accuracy and thus the number of true and false-positive SNPs in the filtered set are shown in Table 3. The low homology filter eliminated between 44 and 89% of the false positive SNP calls that were in the unfiltered set, while eliminating fewer than 1% of the true positives in all the experiments. The footprint effect filter had a much larger impact on the SCHU S4 results than on the LVS results. This is not surprising since this filter eliminates spurious SNP calls that occur because of genuine SNPs in the experimental DNA sample, and the SCHU S4 sample contains a substantially larger number of genuine SNPs compared with LVS. Each successive filter further reduced the set of false-positive SNP calls with a cost in terms of a loss of some true-positive SNPs. The cumulative effect was a reduction of the initial false-positive set by over 98% in the case of LVS and over 91% in the case of SCHU S4, with no loss of true positives for LVS and a loss of about 10.7% of true positives for SCHU S4 after implementation of the filters. The overall base-calling accuracy (considering both reference calls and SNP calls) was 99.999% for LVS and 99.992% for SCHU S4. These accuracy rates are equivalent to Phred quality scores of 50 and 41, respectively.

We did not expect to find any genuine SNP calls in the results from the LVS experiments, as the design of the resequencing chips was primarily based on this genome sequence. ABI sequence validation of 290 SNP calls from the unfiltered data sets confirmed three of these LVS base calls as true SNPs. ABI sequence validation of a subset of 562 possible SNP locations from the SCHU S4 strain (versus the LVS reference strain) are shown in Table 4. Sequence results were obtained for 484 (~86.1%) of the 562 selected locations, of which 320 and 164 were believed to be false positives and false negatives, respectively, based on previously published sequences. Only 5 out of the 320 suspected false positives were true SNPs. Similarly, six of the suspected false negatives were found to be true negatives in our validation results. It is not clear whether these differences are due to errors in the published SCHU S4 genomic sequence or they represent genuine point mutations in our sample DNA as compared to the SCHU S4 strain DNA used to obtain the published genomic sequence. The complete lists of validated SNP locations and the results are shown in Supplementary Tables 2 and 3.

One outcome that initially seemed puzzling was the much larger number of false-positive SNPs in the SCHU S4 data that remained after the data were treated with the filters. This outcome was directly related to the nature of the SCHU S4 sample. A mapping of the SCHU S4 genome onto the LVS reference sequence, produced with the MUMer software, reveals many rearrangements in SCHU S4 relative to LVS. A feature probe on the resequencing chip that spans one of these rearrangement boundaries would be expected to hybridize poorly with the SCHU S4 sample, and this increased the likelihood that alternate sequences would hybridize more strongly, increasing the false-positive calls. This outcome suggests that the resequencing platform could be used to identify

Table 2. Raw resequencing data for *F. tularensis* LVS and *F. tularensis* SCHU S4 samples

Expt. No.	Array	Bases/array	Bases called	Call rate (%)	SNPs	% SNPs of called bases
Raw Data for *F. tularensis* LVS						
007	A	301 470	298 283	98.943	30	0.010
	B	302 018	298 072	98.693	30	0.010
	C	301 394	297 350	98.658	35	0.012
	D	296 905	292 333	98.460	43	0.015
	E	290 100	283 408	97.693	21	0.007
	F	273 824	230 750	84.269	1087	0.471
	Total	1 765 711	1 700 196	96.290	1246	0.073
013	A	301 470	297 426	98.659	30	0.010
	B	302 018	297 614	98.542	28	0.009
	C	301 394	297 381	98.669	38	0.013
	D	296 905	291 534	98.191	45	0.015
	E	290 100	285 828	98.527	27	0.009
	F	273 824	234 169	85.518	1688	0.721
	Total	1 765 711	1 703 952	96.502	1856	0.109
Raw data for *F. tularensis* SCHU S4						
008	A	301 470	288171	95.589	1331	0.462
	B	302 018	291499	96.517	1293	0.444
	C	301 394	291 988	96.879	1571	0.538
	D	296 905	282 940	95.296	1545	0.546
	E	290 100	280 992	96.860	1306	0.465
	F	273 824	258 411	94.371	1326	0.513
	Total	1 765 711	1 694 001	95.939	8372	0.494
014	A	301 470	292 313	96.963	1383	0.473
	B	302 018	290 452	96.170	1298	0.447
	C	301 394	290 080	96.246	1532	0.528
	D	296 905	282 768	95.239	1539	0.544
	E	290 100	280 557	96.710	1293	0.461
	F	273 824	256 803	93.784	1259	0.490
	Total	1 765 711	1 692 973	95.881	8304	0.490

For these results, a quality score threshold of 12 and a call rate cutoff of zero were used, as explained in the text. All base positions on the chip set were considered (LVS, SCHU S4 and plasmid reference sequences). This accounts for the much higher SNP count from chip F in the LVS experiments.

genome rearrangements. We found that the majority of the false-positive SNP calls in the SCHU S4 sample fell into one of two categories: (i) those that lie within 12 bases of a rearrangement boundary and (ii) those that lie within 12 bases of a predicted SNP. These results are summarized in Table 5. In spite of the larger number of false positives in the SCHU S4 data set, they represent only 2.04% of the SNP calls that remained after filtering.

Table 6 shows the comparison of raw and filtered data for LVS and SCHU S4. The raw call rate and accuracy take into account all base positions on the resequencing chips and report the results prior to any filtering steps. The genome-adjusted results take into account only those portions of the chips that have high sequence homology with the hybridized sample. The data indicated a false-negative SNP rate in the range of 0–17.31% and a false-positive rate in the range of 0.001–0.007%. The false-positive SNP rate is the number of false positives divided by the number of bases at which a genuine SNP call was not expected. The false-negative SNP rate is the number of expected SNPs that were not identified divided by the total number of expected SNPs. The false-negative rate can be misleading, since this rate includes all expected SNPs that were not detected, including those that were not in the raw data set as well as those that were removed by our filters. Although the false-negative SNP rate for the SCHU S4 sample was 17.310%, it is important to note that the filters eliminated less than 11% of the true-positive SNPs that were in the raw data set (see Table 3). There is an inevitable tradeoff between the rejection of false positives and the retention of true-positive SNPs. In general, an increase in the stringency of filtering will cause a reduction in both false positives and true positives. The filtering scripts can be parameterized by the user for an appropriate tradeoff between sensitivity (retention of true positives) and specificity (rejection of true negatives). Since LVS was the primary reference whose sequence is represented most fully on the chips, the results for the LVS samples were better than we would expect to achieve with a sample of unknown composition. The efficiency of the platform cannot be numerically defined as it varies according to the extent of the difference between the sample DNA and the reference sequence.

DISCUSSION

The comparative analysis of multiple genomic sequences highlights the extreme variability that exists within many, if not most, microbial species (14,15). Sequence information from multiple species of a selected few microbes have clearly indicated that a single reference genome only provides a limited genomic overview of a species (16) and is not sufficient in understanding the genetic potential of an organism. The genomic plasticity evident in microbial

Table 3. Effects of filtering steps on base calling accuracy

Filter steps	True positives	False positives	Accuracy (%)	True-positive retention (%)	False-positive rejection (%)
F. tularensis LVS (Expt. # 007)					
None (raw unfiltered)	3	1243	99.927	100.000	0.000
Low homology	3	179	99.989	100.000	85.599
Alternate homology	3	25	99.999	100.000	97.989
Footprint effect	3	23	99.999	100.000	98.150
Replicate combination	3	19	99.999	100.000	98.471
F. tularensis LVS (Expt. # 013)					
None (raw unfiltered)	3	1853	99.891	100.000	0.000
Low homology	3	190	99.989	100.000	89.746
Alternate homology	3	30	99.998	100.000	98.381
Footprint effect	3	29	99.998	100.000	98.435
Replicate combination	3	19	99.999	100.000	98.975
F. tularensis SCHU S4 (Expt. # 008)					
None (raw unfiltered)	6908	1464	99.914	100.000	0.000
Low homology	6878	816	99.951	99.566	44.262
Alternate homology	6529	388	99.977	94.514	73.497
Footprint effect	6327	200	99.988	91.589	86.339
Replicate combination	6172	126	99.992	89.346	91.393
F. tularensis SCHU S4 (Expt. # 014)					
None (raw unfiltered)	6902	1402	99.917	100.000	0.000
Low homology	6859	777	99.954	99.377	44.579
Alternate homology	6515	363	99.978	94.393	74.108
Footprint effect	6317	198	99.988	91.524	85.877
Replicate combination	6172	126	99.992	89.423	91.013

The true positive retention and false positive rejection rates are calculated relative to the number of true and false positive results in the raw, unfiltered data. The accuracy is calculated relative to the number of base calls remaining after the specified filtering step, where reference calls and true positive SNP calls are considered correct, and no-calls ('N') are not considered.

Table 4. SCHU S4 SNP validation summary

Total locations attempted	562
Results obtained	484
False-positive validation results	320
False-negative validation results	164
'False positive' calls revealed as 'True positive'	5
'False positive' calls confirmed as 'False positive'	315
'False negative' calls revealed as 'True negative'	6
'False negative' calls confirmed as 'False negative'	158

Table 5. Causes of false-positive SNP calls in SCHU S4

Category	Number of SNPs
Total false positives after filtering	126
False-positive SNPs within 12 bases of a rearrangement boundary	61
False-positive SNPs within 12 bases of a predicted SNP	29
Unexplained false-positive SNPs	42

A total of six false-positive SNPs were found to be both within 12 bases of a rearrangement boundary and within 12 bases of a predicted SNP.

genomes has drawn into question the relative value of any single reference genomic DNA sequence. Polymorphisms in the form of SNPs, indels and genomic rearrangements are common (17,18). Applications in areas of comparative microbial genomics, molecular microbial forensics, molecular epidemiology, biodefense and evolution demand more genomic-scale sequence information from multiple isolates than ever before.

Directed resequencing strategies are generally considered a more efficient and economical approach as compared with the classical shotgun sequencing strategies (7,19). The increased availability of complete DNA sequence data for a large number of reference genomes has elevated the value of resequencing methods. Sequencing by hybridization is a high-throughput DNA sequencing platform and offers the potential to improve our understanding of diversity within and across species. The quality of microarray-generated DNA sequence data is directly comparable to that produced by conventional shotgun sequencing (20). The global identification of SNPs in genomes represents a case where a resequencing by hybridization approach may be favored over other alternatives.

Resequencing for point mutations using microarrays was demonstrated in 1996 (21) and has become an established methodology (3). Only one high-quality reference genome sequence is required for sequencing multiple strains of the same organism. High-density resequencing microarrays offer a unique opportunity to genotype microorganisms at a nucleotide resolution, providing reliable and accurate information for identifying, typing and tracking infectious and bio-threat agents. Read and co-workers have demonstrated the power of comparative full-genome sequencing and identification of genetic polymorphisms in two related strains of *Bacillus anthracis* (22). Resequencing arrays have also been used to detect group A streptococci and their associated antibiotic resistance markers (8). Whole genomes of severe acute respiratory syndrome (SARS) virus have

Table 6. Comparison of raw (unfiltered) versus filtered resequencing results

Results	LVS (007)	LVS (013)	SCHU S4 (008)	SCHU S4 (014)
F. tularensis sample (Experiment No)				
Raw				
Raw positions	1 765 711	1 765 711	1 765 711	1 765 711
Raw base calls	1 700 196	1 703 952	1 694 001	1 692 973
Raw call rate	96.290%	96.502%	95.939%	95.881%
Raw accuracy	99.927%	99.891%	99.914%	99.917%
False positive SNPs	1243	1853	1464	1402
True positive SNPs	3	3	6908	6902
Genome-adjusted				
Genome-adjusted positions	1 725 937	1 725 937	1 743 224	1 743 224
Filtered base calls	1 689 733	1 689 733	1 674 222	1 674 222
Filtered call rate	97.902%	97.902%	96.042%	96.042%
Filtered accuracy	99.999%	99.999%	99.992%	99.992%
False-positive SNPs	19	19	126	126
True-positive SNPs	3	3	6172	6172
False-negative SNPs	0	0	1292	1292
False-positive SNP rate	0.001%	0.001%	0.007%	0.007%
False-negative SNP rate	0.000%	0.000%	17.310%	17.310%

The genome-adjusted results are calculated relative to the portions of the chip set that performed well with the DNA samples under consideration. The regions identified by our low-homology filter are excluded from the genome-adjusted positions. The false-negative SNP counts represent the number of expected SNPs that were missing from the final, filtered SNP set. For SCHU S4, 7464 SNPs were expected, on the basis of *in silico* alignment of the LVS and SCHU S4 genome sequences. In the false-positive SNP rate calculation, the denominator is the number of genome-adjusted base positions that were not expected to be SNPs. In the false-negative SNP rate calculation, the denominator is the number of genome-adjusted positions that were expected to be SNPs.

been characterized using a resequencing array approach (5,9). Most recently, resequencing of 14 smallpox virus genomes has been reported using a set of seven 30 K GeneChip arrays and a classical PCR-based approach (10). These studies have shown the promise of the technology and the limitations associated with no-calls and mis-called bases. PCR amplification of long targets (>10 kb) generally used in resequencing is non-trivial and limits sample processing throughput.

Zwick *et al.* (20) have sequenced 0.5% of the *B. anthracis* genome from 56 strains using resequencing arrays and recognized a need for resequencing of a larger percentage of the genome from multiple isolates to detect rare recombination events. Whole-genome resequencing of multiple strains will provide an enhanced genomic overview and a higher-resolution genotyping of an organism.

The whole-genome resequencing approach described here made use of genome amplification (23,24), a method that has been successfully used in genotyping (25,26) and molecular epidemiological (27) studies. The technical modification of elimination of the PCR amplification step from the resequencing procedure makes the platform more cost effective and amenable to high throughput. The resequencing array platform by nature is highly efficient. A single person can easily generate whole-genome sequence and SNP data for at least four *F. tularensis* genomes per week without any automation using our approach. This can be further improved by technological advances in array feature density, automation and increasing the man power, providing whole-genome sequence and SNP information from multiple isolates in a matter of days.

Systematic effects causing false-positive SNP calls in the resequencing platform are exacerbated by the hybridization of a whole-genome sample, as compared with a sample consisting of purified selected PCR fragments. A whole-genome sample contains a much larger complexity of targets for hybridization than does a PCR-amplified subset of the whole genome, and these additional targets give rise to some of the inaccurate base calls. The widely accepted ABACUS algorithm (4) for the Affymetrix-based resequencing platform may not perform with equal efficiency on a whole-genome query sample. It is necessary to understand the nature and source of errors in whole-genome hybridization in order to optimally use this platform. Improved base-calling algorithms and bioinformatic tools are essential to decrease false positives, minimize false negatives and increase the overall base-calling accuracy of this platform. The bioinformatic filters reported here were developed towards achieving these goals.

The raw whole-genome resequencing data, even though apparently acceptable, has a rather large number of false positives (Table 2), indicating an inherent limitation of the existing data analysis tool on the Affymetrix resequencing platform. We have developed a set of software filters to identify and filter out SNP calls that are likely to be artifacts. These filters operate on the base calls and quality scores provided by the Affymetrix GSEQ program, yielding a set of high-confidence SNP calls containing a lower percentage of false positives. We used the set of SNPs predicted by *in silico* analysis for the hybridization of the SCHU S4 sample to test and fine-tune our filters. Our bioinformatic filters identified many of these false positives (Table 3).

The computational requirements of these filters are modest. We routinely run them on a 2.8 MHz Xeon® dual-processor Linux workstation. A set of six *F. tularensis* chips representing one experiment can be processed in an

overnight run in this environment. However, this time can typically be reduced to less than 1 h if a precomputed database of binding energies is used. The most computationally intensive part of the filter algorithms is the identification of potential alternate homologies in the reference sequence, and the calculation of binding energies for these sequences. But this task need only be performed the first time that a particular SNP location is encountered. We store the results of these calculations in a data file, and when new experiments are processed, results for previously encountered SNP locations are looked up in this database rather than computed.

Some of the issues identified in our results suggest some areas for further improvement of our filters. The presence of false-positive SNP calls that occur near rearrangement boundaries indicates that the low-homology filter did not identify all low-homology regions that could potentially be removed, and its performance could be further improved. Identifying a signature in the resequencing data that corresponds to a rearrangement boundary may allow us to filter false-positive SNPs from this source. More importantly, this boundary information itself should prove to be a valuable tool for the genotyping of different strains of an organism. The occurrence of coincident rearrangement boundaries in two different samples is strong evidence of genetic similarity and can be used in conjunction with the SNP data to draw conclusions about phylogenetic relationships.

Similarly, the retention of false positives that lie near predicted SNP locations indicates that the footprint effect filter also can be further refined. The footprint effect filter also had a relatively large negative impact on true positives in the SCHU S4 results (see Table 3). One underlying assumption of the current filter is that, within a set of closely spaced SNP calls, the one with the highest quality score is most likely to be the genuine SNP. This assumption may be violated in some cases, resulting in both false-positive and false-negative SNP calls. Another weakness of the current filter is that when two SNP calls lie within 10 base positions of each other, only one of the calls can survive the filter. Perhaps the effectiveness of this filter can be improved by examining signal intensities directly, rather than relying solely on the quality values assigned by the GSEQ software.

We used the set of expected SNPs present in the SCHU S4 sample, relative to the LVS reference sequence, to parameterize and validate our filter algorithms. Consequently, the possibility of over-fitting of the parameters to one particular genome sequence cannot be ignored. However, the excellent performance of the filters on the LVS query sample argues against this possibility. Additional experiments with the recently available genome sequence of a clinical strain WY96-3418 (GenBank accession number CP000608) also support the robustness of the filter parameters chosen. We used the published sequence data to predict the expected SNP calls and validate our results for WY96-3418, in the same way as was done for SCHU S4. Our filters eliminated over 95% of the false-positive SNP calls and achieved a call accuracy rate of 99.995%, equivalent to a Phred quality score of 43 (Supplementary Table 4).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Luther Lindler of USAMRIID for providing us with genomic DNA samples of reference *F. tularensis* LVS and SCHU S4 strains. We also thank the Diagnostic and Reference Laboratory, CDC, Fort Collins (CO) for providing us with the *F. tularensis* WY96-3418 genomic DNA for our studies. We sincerely acknowledge Dr Mark J. Wolcott of USAMRIID Fort Detrick (MD) and the Diagnostic and Reference Laboratory, CDC, Fort Collins (CO) for providing us with genomic DNA of *F. tularensis s*trains used in batch analysis. We acknowledge technical suggestions and advice on resequencing arrays and data analysis from Dione Bailey, Anna Berdine and Francois Collin of Affymetrix, Inc. This work was supported by the Department of Homeland Security and the NIAID contract No. NO1-AI-15447. Funding to pay the Open Access publication charges for this article was provided by the National Institute of Allergy and Infectious Diseases (NIAID).

Conflict of interest statement. None declared.

REFERENCES

1. Cebula,T.A., Jackson,S.A., Brown,E.W., Goswami,B. and LeClerc,J.E. (2005) Chips and SNPs, bugs and thugs: a molecular sleuthing perspective. *J. Food Prot*., **68**, 1271–1284.
2. Mockler,T.C., Chan,S., Sundaresan,A., Chen,H., Jacobsen,S.E. and Ecker,J.R. (2005) Applications of DNA tiling arrays for whole-genome analysis. *Genomics*, **85**, 1–15.
3. Hacia,J.G. (1999) Resequencing and mutational analysis using oligonucleotide microarrays. *Nat. Genet*., **21**, 42–47.
4. Cutler,D.J., Zwick,M.E., Carrasquillo,M.M., Yohn,C.T., Tobin,K.P., Kashuk,C., Mathews,D.J., Shah,N.A., Eichler,E.E *et al.* (2001) High-throughput variation detection and genotyping using microarrays. *Genome Res*., **11**, 1913–1925.
5. Wong,C.W., Albert,T.J., Vega,V.B., Norton,J.E., Cutler,D.J., Richmond,T.A., Stanton,L.W., Liu,E.T. and Miller,L.D. (2004) Tracking the evolution of the SARS coronavirus using high-throughput, high-density resequencing arrays. *Genome Res*., **14**, 398–405.
6. Maitra,A., Cohen,Y., Gillespie,S.E., Mambo,E., Fukushima,N., Hoque,M.O., Shah,N., Goggins,M., Califano,J. *et al.* (2004) The human mitoChip: a high-throughput sequencing microarray for mitochondrial mutation detection. *Genome Res*., **14**, 812–819.
7. Shendure,J., Mitra,R.D., Varma,C. and Church,G.M. (2004) Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet*., **5**, 335–344.
8. Davignon,L., Walter,E.A., Mueller,K.M., Barrozo,C.P., Stenger,D.A. and Lin,B. (2005) Use of resequencing oligonucleotide microarrays for identification of *Streptococcus pyogenes* and associated antibiotic resistance determinants. *J. Clin. Microbiol*., **43**, 5690–5695.
9. Sulaiman,I.M., Liu,X., Frace,M., Sulaiman,N., Olsen-Rasmussen,M., Neuhaus,E., Rota,P.A. and Wohlhueter,R.M. (2006) Evaluation of Affymetrix severe acute respiratory syndrome resequencing GeneChips in characterization of the genomes of two strains of coronavirus infecting humans. *Appl. Enivron. Microbiol*., **72**, 207–211.
10. Sulaiman,I.M., Tang,K., Osborne,J., Sammons,S. and Wohlhueter,R.M. (2007) GeneChip resequencing of the smallpox

virus genome can identify novel strains: a biodefense application. *J. Clin. Microbiol.*, **45**, 358–363.
11. Delcher,A.L., Kasif,S., Fleischmann,R.D., Peterson,J., White,O. and Salzberg,S.L. (1999) Alignment of whole genomes. *Nucleic Acids Res.*, **27**, 2369–2376.
12. Bozdech,Z., Zhu,J., Joachimiak,M.P., Cohen,F.E., Pulliam,B. and DeRisi,J.L. (2003) Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. *Genome Biology*, **4**, R9 http://genomebiology.com/2003/4/2/R9
13. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
14. Whittam,T.S. and Bumbaugh,A.C. (2002) Inferences from whole-genome sequences of bacterial pathogens. *Curr. Opin. Genet. Dev.*, **12**, 719–725.
15. Thomson,N., Sebaihia,M., Cerdeno-Tarraga,A., Bentley,S., Crossman,L. and Parkhill,J. (2003) The value of comparison. *Nat. Rev. Microbiol.*, **1**, 11–12.
16. Tettelin,H., Masignani,V., Cieslewicz,M.J., Donati,C., Medini,D., Ward,N.L., Angiuoli,S.V., Crabtree,J., Jones, and , (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial ''pan-genome''. *Proc. Natl Acad. Sci. USA*, **102**, 13950–13955.
17. Fraser-Liggett,C.M. (2005) Insights on biology and evolution from microbial genome sequencing. *Genome Res.*, **15**, 1603–1610.
18. Abby,S. and Daubin,V. (2007) Comparative genomics and the evolution of prokaryotes. *Trends Microbiol.*, **15**, 135–141.
19. Bentley,D.R. (2006) Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.*, **16**, 545–552.
20. Zwick,M.E., Mcafee,F., Cutler,D.J., Read,T.D., Ravel,J., Bowman,G.R., Galloway,D.R. and Mateczun,A. (2004) Microarray-based resequencing of multiple Bacillus anthracis isolates. *Genome Biol.*, **6**, R10.
21. Chee,M., Yang,R., Hubbell,E., Berno,A., Huang,X.C., Stern,D., Winkler,J., Lockhart,D.J., Morris,M.S. *et al.* (1996) Accessing genetic information with high-density DNA arrays. *Science*, **274**, 610–614.
22. Read,T.D., Salzberg,S.L., Pop,M., Shumway,M., Umayam,L., Jiang,L., Holtzapple,E., Busch,J.D., Smith,K.L. *et al.* (2002) Comparative genome sequencing for discovery of novel polymorphisms in Bacillus anthracis. *Science*, **296**, 2028–2033.
23. Dean,F.B., Hosono,S., Fang,L., Wu,X., Faruqi,A.F., Bray-Ward,P., Sun,Z., Zong,Q., Du,Y. *et al.* (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl Acad. Sci.*, **99**, 5261–5266.
24. Dean,F.B., Nelson,J.R., Giesler,T.L. and Lasken,R.S. (2001) Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.*, **11**, 1095–1099.
25. Barker,D.L., Hansen,M.S., Faruqi,A.F., Giannola,D., Irsula,O.R., Lasken,R.S., Latterich,M., Makarov,V., Oliphant,A. *et al.* (2004) Two methods of whole-genome amplification enable accurate genotyping across a 2320-SNP linkage panel. *Genome Res.*, **14**, 901–907.
26. Tzvetkov,M.V., Becker,C., Kulle,B., Nurnberg,P., Brockmoller,J. and Wojnowski,L. (2005) Genome-wide single-nucleotide polymorphism arrays demonstrate high fidelity of multiple displacement-based whole-genome amplification. *Electrophoresis*, **26**, 710–715.
27. Yan,J., Feng,J., Hosono,S. and Sommer,S.S. (2004) Assessment of multiple displacement amplification in molecular epidemiology. *Biotechniques*, **37**, 136–143.