SCIENTIFIC REPORTS

OPEN

Genetic variation and linkage disequilibrium in *Bacillus anthracis*

SUBJECT AREAS:
PATHOGENS
MICROBIAL GENETICS
MOLECULAR EVOLUTION
GENETIC VARIATION

Received
11 August 2011

Accepted
3 November 2011

Published
24 November 2011

Correspondence and requests for materials should be addressed to M.Z. (mzwick@emory.edu)

Michael E. Zwick[1,2], Maureen Kiley Thomason[1], Peter E. Chen[1], Henry R. Johnson[4], Shanmuga Sozhamannan[1], Alfred Mateczun[1] & Timothy D. Read[1,2,3]

[1]Biological Defense Research Directorate, Naval Medical Research Center, Silver Spring, MD, USA, [2]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, USA, [3]Division of Infectious Diseases, Emory University School of Medicine, Atlanta, GA, USA, [4]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD.

We performed whole-genome amplification followed by hybridization of custom-designed resequencing arrays to resequence 303 kb of genomic sequence from a worldwide panel of 39 *Bacillus anthracis* strains. We used an efficient algorithm contained within a custom software program, UniqueMER, to identify and mask repetitive sequences on the resequencing array to reduce false-positive identification of genetic variation, which can arise from cross-hybridization. We discovered a total of 240 single nucleotide variants (SNVs) and showed that *B. anthracis* strains have an average of 2.25 differences per 10,000 bases in the region we resequenced. Common SNVs in this region are found to be in complete linkage disequilibrium. These patterns of variation suggest there has been little if any historical recombination among *B. anthracis* strains since the origin of the pathogen. This pattern of common genetic variation suggests a framework for recognizing new or genetically engineered strains.

Characterizing the patterns of genomic variation found among microbial pathogens often reveals unique aspects of their biology and evolutionary history[1,2]. In bacteria and archaea, recombination arises from proximate mechanisms that include transduction, conjugation, and transformation, and can shape the levels of genomic variation and the observed patterns of statistical association between variant sites[3]. Detecting the patterns of association among common variant sites, termed linkage disequilibrium, has been used in both bacteria and archaea to help elucidate the effects of recombination on genomic variation[4–7]. More recently, elegant methods for detecting recombination have confirmed that historical recombination rates show extraordinary levels of variation within some bacteria genera[8,9].

Whole-genome sequencing studies of the highly virulent gram-positive endospore-forming bacterium *Bacillus anthracis*, the agent used in the 2001 bioterrorist attacks in the United States, have led to a number of major findings. *B. anthracis* is found to be a recently emerged, monophyletic lineage from within the polyphyletic *Bacillus cereus* sensu lato group, with low levels of genetic variation. A variety of approaches, including multilocus variable number of tandem repeats analysis (MLVA)[10–16], amplified fragment length polymorphism[17], Sanger sequencing[18–20], multilocus sequence typing (MLST)[21], and microarray-based resequencing[22], note a paucity of genetic variation within *B. anthracis*. Complete genome sequencing of a limited number of *B. anthracis* genomes lends further support for these findings[23,24]. Recently, reports show that genotyping *B. anthracis* strains with "canonical" single nucleotide polymorphism (SNP) typing can efficiently illuminate the organism's global population structure[25].

While the *B. anthracis* lineage appears to be monophyletic, the existence of "canonical SNPs" implies that historical recombination between strains is likely a rare event. In a previous study, we used custom-designed resequencing arrays to resequence 29 kb from a worldwide panel of 56 *B. anthracis* strains (3.1 Mb total sequence). Our analysis showed not only low levels of genetic variation, but also complete linkage disequilibrium among the common single nucleotide variants we discovered[22]. These variant sites were located on the pXO1 and pXO2 plasmids in addition to the main chromosome. These observations are consistent with a model in which all extant *B. anthracis* strains arose from a single clone, with no historical recombination occurring among the different strains; thus, the common variants observed today in *B. anthracis* strains reflect the mechanism of mutation as opposed to the acquisition of sequences from other strains by recombination. If true, this hypothesis would explain why it is possible to use a few canonical SNPs to characterize the global population structure of clonal *B. anthracis* strains.

Here we report the results of a sequencing study whose aim was to quantitatively assess the levels and patterns of genomic variation in *B. anthracis* to replicate our original findings for a much larger genomic region. Using a novel experimental protocol consisting of whole-genome amplification of different samples followed by hybridization to custom-designed resequencing arrays, we resequenced 303 kb in each of 39 *B. anthracis* strains from a worldwide strain collection (9.6 Mb total sequence). We used an efficient algorithm contained within a custom software program, UniqueMER, to identify and mask repetitive sequences on the resequencing array to reduce false-positive identification of genetic variation, which can arise due to cross-hybridization. Our analysis of the resulting sequencing data estimates a remarkably low level of DNA sequence variation, by functional class, in *B. anthracis*. Furthermore, our analysis shows complete linkage disequilibrium among common segregating sites in the region that we resequenced. The patterns of variation we see are consistent with an absence of historical recombination among *B. anthracis* strains since the origin of the pathogen.

Results

We performed targeted sequencing of 39 *B. anthracis* strains from the Biological Defense Research Directorate's strain collection using custom-designed Affymetrix resequencing arrays (Table 1). We determined the raw sequence from each RA image file by using the ABACUS algorithm as implemented within the RATools software package[22,26,27], and then filtered as described in the Materials and Methods. A total of 9.5 Mb (~245 kb per *B. anthracis* strain) of genome sequence was obtained (Supplemental File 1, Supplemental Table 2). Figure 1 shows the phylogeny of the *B. anthracis* strains inferred from these sequences. Two results are apparent. First, we see a statistically significant differentiation between the *B. anthracis* A and B strains, as found previously by both ourselves and others[13,22,25]. Second, we see that the sequenced Ames strains cluster together (BAN 003, 039, 032, 041), which reflects the recent origin of these strains from a common ancestor.

Population genomic analysis of the 39 *B. anthracis* strains sequenced revealed a total of 240 single nucleotide variants (SNVs), with strains having an average of 2.25 differences per 10,000 bases sequenced (Table 2). This analysis shows that *B. anthracis* has a remarkably low level of genomic variation, consistent with our previous estimate and what has been seen in a number of newly arising bacterial pathogens[2,22]. For the purpose of comparison, this level of variation is roughly a quarter of that observed in the human genome when sampled in a similar worldwide fashion[28–30]. After functionally annotating the 240 SNVs, we found that on a per-site basis, replacement sites (those sites that change amino acids in proteins) are the least variable, silent sites are the most variable, and intergenic regions have intermediate levels of genetic variation.Our data provide a slightly lower (0.39 vs 0.58) albeit not dramatically different estimate for the dN/dS ratio than that previously reported for *B. anthracis* (Table 2).

Table 1 | List of the worldwide collection of 39 *B. anthracis* strains resequenced

Sample ID	Species ID	MLVA Genotype (Cluster)	Strain Information
BAN_003	ASC159	62(A2)	Texas, USA. Ames strain. Guinea pig re-isolate.
BAN_004	ET-76B		Etosha National Park. Namibia.
BAN_005	NMRC-GT41-001	41(A3a)	GT-41
BAN_006	NMRC-GT68-003	68(A3)	GT-68
BAN_007	NMRC-BACI008-003P		
BAN_008	NMRC-GT28-02A1		
BAN_009	NMRC-BACI056		
BAN_010	NMRC-GT3-007	3(A1a)	
BAN_011	ASC069		New Hampshire, USA. Human isolate.
BAN_012	A0039	55(A3a)	Australia. Bovine isolate.
BAN_013	ASC015		ATCC 00938
BAN_014	A0248	68(A3d)	USA. Human isolate.
BAN_015	7702-2	59 or 61(A2)	Sterne 7702 (pXO1+)
BAN_016	ASC285		UK. Environmental isolate.
BAN_017	NMRC-VOLLUM-002	77(A4)	Vollum
BAN_018	ASC014		ATCC 00241
BAN_019	A0174	3(A1a)	Canada
BAN_020	ASC031		UK. Bovine case, contaminated material from Senegal.
BAN_021	ASC006	77(A4)	UK. Vollum 3b type strain.
BAN_022	ASC038		UK. Fatal human case.
BAN_023	ASC061		Etosha National Park. Namibia. Zebra isolate.
BAN_024	A0328	38(A3a)	Germany. Pig isolate.
BAN_025	ASC016		ATCC 00937
BAN_026	ASC065		Brazil. Cow isolate.
BAN_027	A0379	69(A4)	Pakistan. Wool isolate.
BAN_028	A0463	29(A2)	Pakistan. Sheep isolate.
BAN_029	A0034	57(A3b)	China. Bovine isolate.
BAN_030	ASC050		Zimbabwe. Human cutaneous isolate.
BAN_031	ASC054		Zimbabwe. Human cutaneous isolate.
BAN_032	NMRC-AMES-004	62(A2)	Texas, USA. Ames strain.
BAN_033	NMRC-BACI055-001		Pasteur-like isolate
BAN_034	A0193	10(A1b)	Bovine isolate
BAN_035	ASC004		UK. Strain M36, used in vaccine research.
BAN_036	A0419	43(A3a)	South Korea. Fatal human case.
BAN_037	A0489	45(A3a)	Argentina. Bovine isolate.
BAN_038	LSU442		Kudu, Kruger N.P., South Africa.
BAN_039	NMRC-BACI008-001	62(A2)	Texas, USA. Ames strain.
BAN_040	A0465	80(B1)	France
BAN_041	NMRC-DELTA-AMES-004	62(A2)	Texas, USA. Ames strain.

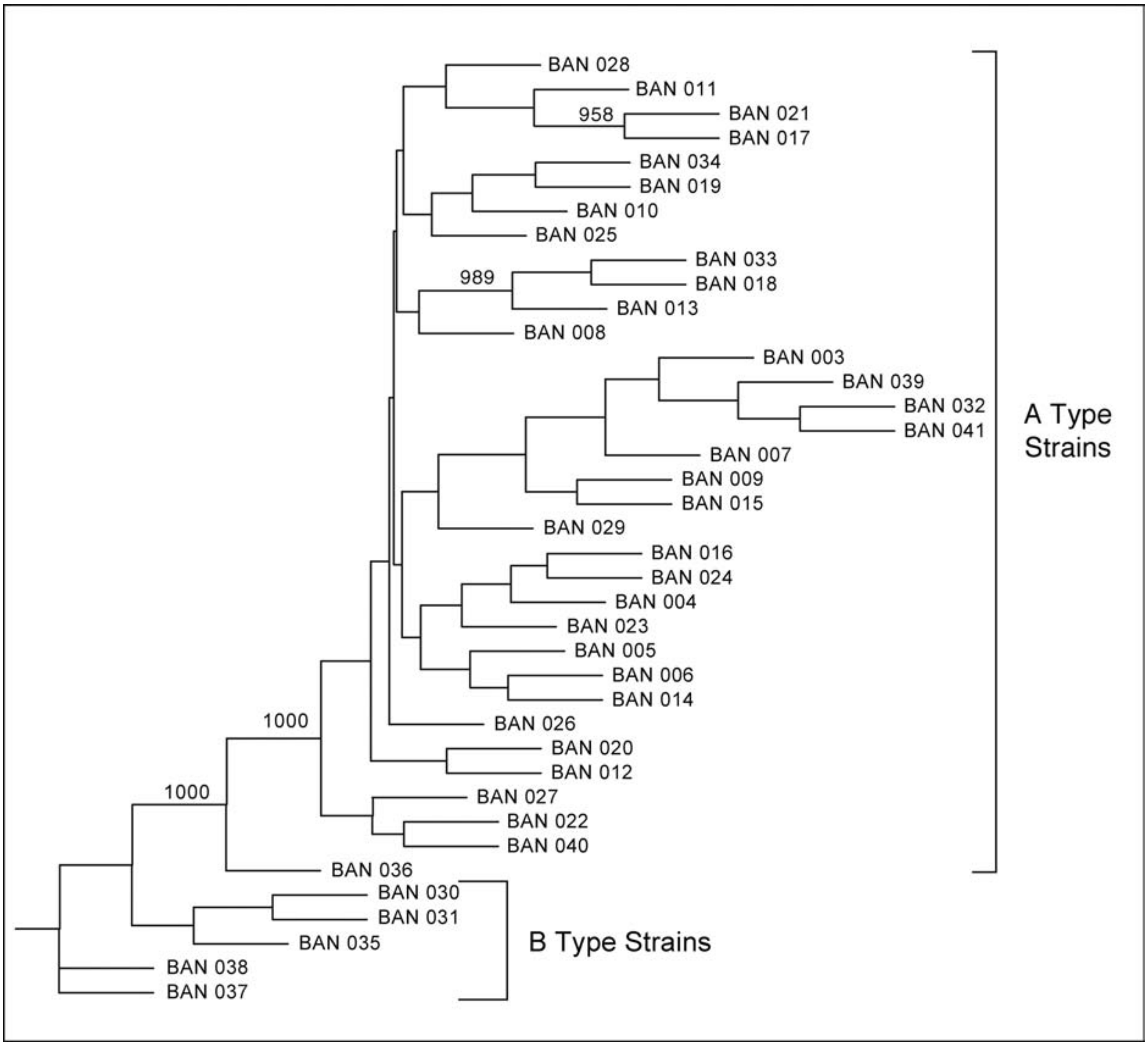

Figure 1 | Phylogenetic relationship of sequences obtained on 39 worldwide *B. anthracis* strains sequenced. Nodes with greater than 95% support from among 1000 bootstrap replicates are shown. The *B. anthracis* B strains cluster together and are found at the base of the tree[13,22,25].

We saw that 141 of the 240 SNVs we discovered were found in just a single sample. To assess the significance of this observation, we performed an analysis of the site frequency spectrum compared with what would be expected under the neutral theory, with the neutral theory expectation assuming we sampled a constant-sized population at mutation-drift equilibrium[31]. Our analysis revealed an excess of rare variants relative to the neutral theory expectation, as evidenced by a negative value for the Tajima's D statistic[32] (Table 2). Statistically significant departures from the neutral expectation were observed for all SNVs (Figure 2) and for the class of replacement SNVs (Figure 3). Possible explanations for the pattern we observed are rapid demographic expansion of *B. anthracis*, or purifying selection acting to remove deleterious alleles, similar to what we reported in our earlier, more limited study[22].

We previously noted an absence of historical recombination in a worldwide collection of *B. anthracis* strains[22]. We predicted that if *B. anthracis* arose from multiple independent clones or underwent recombination since the time of the most recent common ancestor of the worldwide collection of strains we sequenced, there should be genetic evidence of this. To test this hypothesis and characterize the extent to which recombination has shaped patterns of genomic variation in *B. anthracis*, we analyzed our data to seek evidence of historical recombination in the region that we resequenced. We first used LDhat to estimate the amount of recombination among

Table 2 | Characteristics of single nucleotide variants (SNVs) observed within genomic regions sequenced in a worldwide collection of 39 *B. anthracis* strains.

Single Nucleotide Variant Functional Categories	Observed Number of Single Nucleotide Variants (SNVs)	Nucleotide Diversity ± 2 SEs (Θ_w x 10^{-4})	Tajima's D	Statistical Significance of Tajima's D
All	240	2.25 ± 0.85	−1.76	0.029
Silent	66	3.49 ± 1.5	−1.49	0.063
Replacement	119	1.73 ± 0.70	−1.80	0.026
Intergenic	55	3.83 ± 1.1	−1.15	0.13

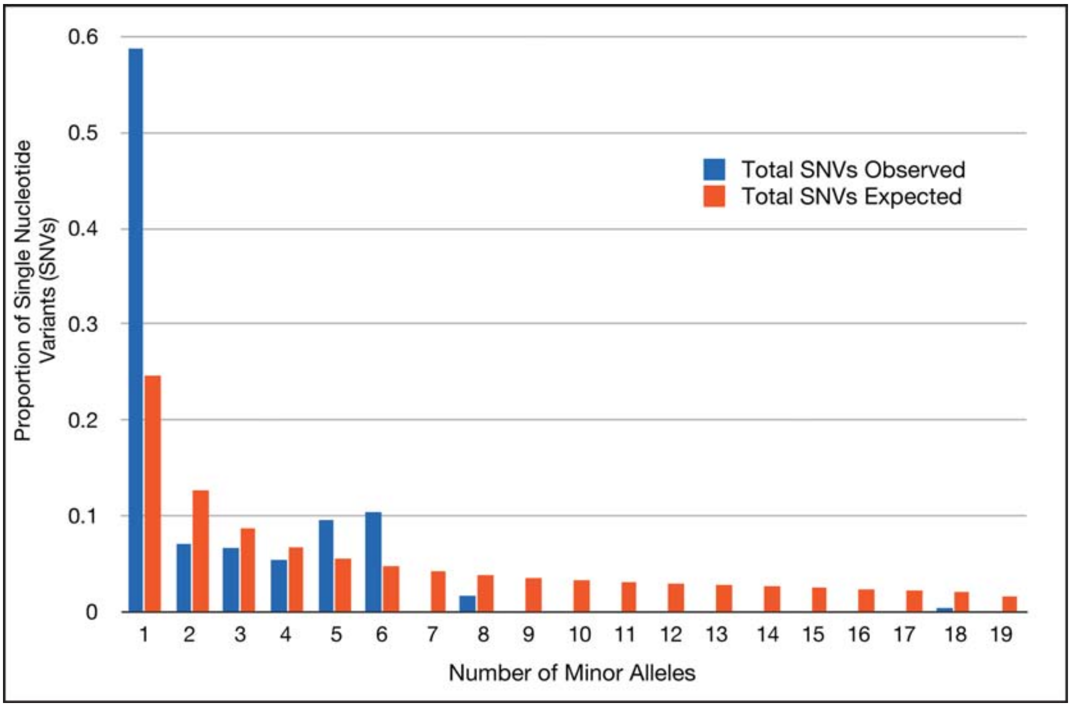

Figure 2 | Histogram reporting the proportion of single nucleotide variants with minor alleles seen one or more times in the 39 worldwide *B. anthracis* strains sequenced. This site frequency spectrum shows an excess of rare single nucleotide variants (blue) relative to the neutral equilibrium expectation (red).

64 common SNVs found at greater than 10% frequency in our sample[33,34]. The 97.5% upper bound for our estimate of historical recombination ($2N_er$) was 1.0×10^{-5} per site. Strikingly, this upper bound estimate for recombination is 22 times lower than that determined for Watterson's estimator of the population mutation rate (Θ_W per site) shown in Table 2[35]. This finding implies that historical recombination has had little or no effect on the patterns of genetic variation we saw. Furthermore, the estimate for $2N_er$ obtained from LDhat is identical to the minimum amount of recombination that can be estimated with this program, implying that the true value could be substantially lower. In a related test, we asked whether historical recombination among any of the 2,278 pairs of common

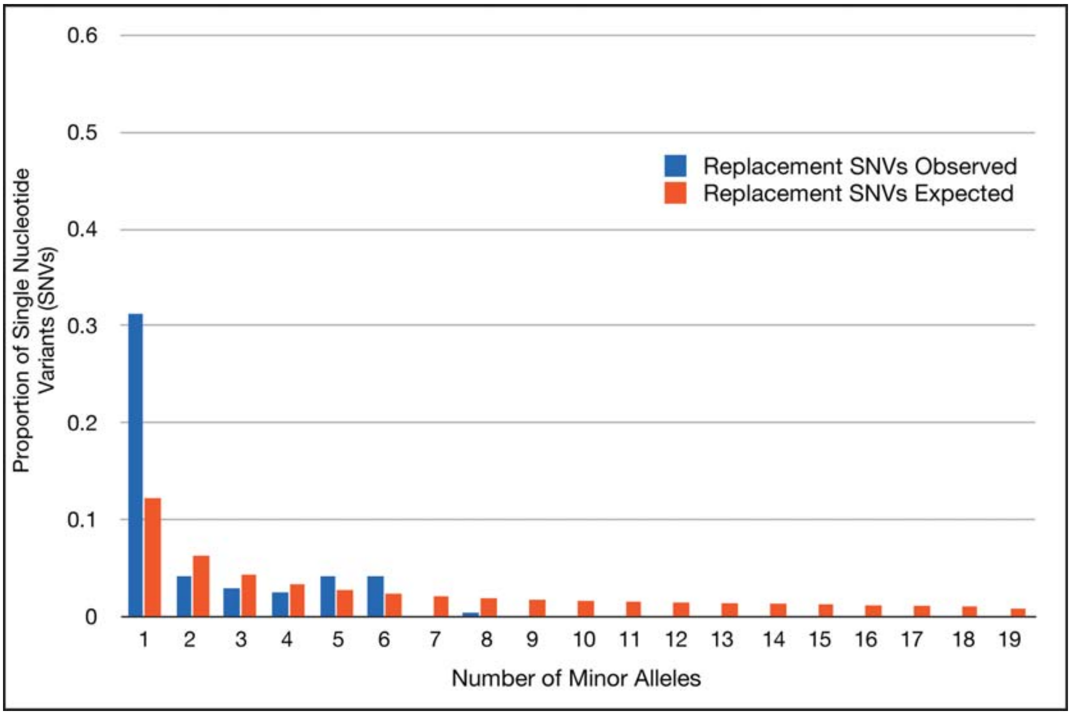

Figure 3 | Histogram reporting the proportion of single nucleotide variants with minor replacement alleles seen one or more times in the 39 worldwide *B. anthracis* strains sequenced. This site frequency spectrum shows an excess of rare replacement single nucleotide variants (blue) relative to the neutral equilibrium expectation (red).

(>10% frequency) SNVs we detected formed four distinct haplotypes[36] as a result of historical recombination between different *B. anthracis* strains. We never saw this outcome in our data, providing a point estimate of 0 for the historical recombination among the SNVs in the region we resequenced[37]. The absence of historical recombination is evident by the complete lack of any pairs of sites with four haplotypes (Figure 4, Haploview 4.2). Combined, our data confirm our previous observation that historical recombination within *B. anthracis* is exceedingly rare or nonexistent, consistent with a model whereby all contemporary *B. anthracis* strains arose from a single common clonal ancestor[22].

Discussion

Our data show that whole-genome amplified bacterial genomes can be hybridized to oligonucleotide resequencing microarrays to determine genome sequences. Analysis of the resulting sequence data gives us an important insight into the population structure and history of *B. anthracis*. A great many studies have supported a monophyletic origin of *B. anthracis*[10–14,17,21,22,25,38] and our analysis reveals no evidence for recombination in the history of the worldwide collection of *B. anthracis* strains we sequenced. This finding provides a clear explanation for why canonical SNPs are able to successfully type strains, because if recombination were common, as has been shown in other larger bacterial genera, then different genomic regions would have distinct evolutionary histories[3,9]. The extensive linkage disequilibrium in *B. anthracis* that we describe stands in stark contrast to some human pathogens, in which exchange of genetic material is fundamental to the organism's pathogenicity[6,7,39–41] (but see[42]).

The apparent absence of historical recombination in *B. anthracis* could have at least two explanations. The first is that *B. anthracis* has reduced recombination, perhaps because it is inherently refractive to transduction, conjugation, and transformation, or because there are defects in the DNA replication machinery. These deficiencies would have to have arisen very early in the history of *B. anthracis* to be passed onto the worldwide population of the species. Arguing against this hypothesis is the observation that genetic studies have shown it is possible to create recombinant *B. anthracis* strains in the laboratory[43]. Furthermore, the induction of natural competence in *B. cereus* ATCC14579[44,45] suggests that transformation could occur in natural populations. Finally, historical recombination has been inferred by comparing genome sequences from strains that compose the larger *B. cereus* group[9]. We believe the more plausible explanation is that low levels of genetic variation combined with the recent global population expansion have limited the opportunities for vegetative *B. anthracis* strains with enough genetic divergence to detect recombination to co-locate. If true, this hypothesis predicts that future dense surveys of *B. anthracis* from Africa, where the most genetically diverse strains are found, might be able to detect recombination, if it is in fact occurring.

An analysis of recently evolved pathogens that included *B. anthracis* reported an elevated dN/dS ratio compared with more distantly related microbial taxa deriving from much more ancient last common ancestors, such as *Escherichia coli*[46]. The authors interpreted their data as providing evidence for relaxed natural selection in newly arising pathogens. This interpretation depends formally upon treating the variants within a clonal lineage, like *B. anthracis*, as older divergent sites (found between species), as opposed to younger, segregating polymorphic sites (found within species). Our data provide a slightly lower (0.39 vs 0.58) albeit not dramatically different estimate for the dN/dS ratio than that previously reported for *B. anthracis* (Table 2). But we disagree with the classification of the sites for recently derived clonal lineages[46]. Rocha et al. (2006) previously showed that comparisons of dN/dS between closely related bacterial genomes need to explicitly consider the time since divergence of the analyzed strains[47]. Furthermore, population genetic theory predicts that the behavior of statistics like dN/dS will differ for polymorphic and divergent sites[48,49] and that the use of this statistic in population-genetic samples is relatively insensitive to the strength of natural selection[50]. In fact, the elevated dN/dS ratios seen are those predicted for segregating polymorphic variants (see Figure 3 in[49]). Thus, the inference of relaxed natural selection in newly arising pathogens, like *B. anthracis*, is not well supported by the data observed.

Finally, the apparent absence of recombination within *B. anthracis* suggests that the patterns of association seen among common sites could be a powerful tool to help recognize newly arising or genetically engineered strains. As new strains are typed for their common SNP variation, their allelic configurations could be compared against other previously characterized strains. Novel allelic configurations would indicate a previously unobserved strain variation and possibly point to a need for greater genetic and phenotypic characterization. The increasing throughput and ever-decreasing costs of pathogen whole-genome sequencing mean that in the very near future, these sorts of sequence-based experiments that can rapidly detect both common and rare variants are likely to become routine[38]. Methods of analysis of these rich datasets that directly characterize the patterns of linkage disequilibrium among variant sites could give us valuable insights into the origins and evolutionary processes shaping the genomes of pathogens.

Methods

***B. anthracis* Strains Sequenced.** We selected a diverse panel of 39 Bacillus strains from the Biological Defense Research Directorate (BDRD) collection at the Navy Medical Research Center (NMRC) for chip resequencing (see Table 1). Twenty-one of the strains were also typed by MLST using ABI sequencing[51]. The MLST data are available through the *Bacillus cereus* MLST website (http://pubmlst.org/bcereus/).

RA design, Hybridization, Sequence determination. The RA design queried 303,006 base pairs and was based upon the *B. anthracis* Ames reference sequence (5.2 Mbp, NC_003997). Unique sequences targeted for sequencing were identified as previously described[22]. Genomic DNA from each strain was isolated using standard protocols as previously described[22,27]. We obtained target DNA for RA hybridization by performing whole-genome amplification (WGA) on 100 ng of genomic DNA following the manufacturer's instructions (REPLI-g Kit from Qiagen, Valencia, CA). The typical yield was 20 – 30 ug per strain. The WGA DNA was then DNAse digested, biotin end-labelled, and hybridized to individual RAs overnight following established protocols[22,27]. Subsequent washes and stains were carried out following the RA manufacturer's standard protocols (Affymetrix, Sunnyvale, CA). RAs were scanned at 570 nm, with a pixel size of 3 m per pixel averaged over 2 scans. Genomic sequences were determined for each sample by using the ABACUS algorithm as implemented in RATools (http://www.dpgp.org)[22,26,27].

Filtering of Raw Sequence Files. The raw sequence files obtained from RATools were filtered in two ways. First, we used UniqueMER to mask repeated 30-mers for each of the 39 strains sequenced. UniqueMER is an open source program that locates all unique and repeat n-mers in an input space consisting of a given set of genomes (https://sourceforge.net/projects/uniquemer/). The algorithm for the program is a distributed hashing scheme consisting of a hash table per computing node. The genomes in the input space are divided among the available computing nodes, and all hash tables are processed in parallel. A hash table can represent one or more genomes, and each entry in a table represents one n-mer and its frequency of occurrence in the entire input space. A sliding window equal to the length of the n-mer slides in 1-bps increments across each genome, the subsequence in the window is hashed, and its frequency of occurrence updated. Each n-mer is hashed using the following hash function:

$$f(s) = s[0] \times 31^{(n-1)} + s[1] \times 31^{(n-2)} + \ldots + s[n-1]$$

where s is the n-mer, s[i] is the ith nucleotide in the n-mer, and n is the length of the n-mer. Thus, an n-mer is unique if it does not have a maximal, exact match to any other n-mer. As space complexity was the bottleneck to allow more genomes to be processed, in-memory load was reduced by avoiding the storage of sequence in the hash tables. Collision resolution for n-mers with identical hash codes but different underlying sequences is achieved by retrieving the n-mer from disk.

A sequence is considered unique if it is unique among all sequences, in both forward and reverse orientations, in the input space. The program tracks the copy number of each n-mer and outputs the frequencies as a histogram. Each n-mer is further grouped into blocks of unique and repeat sequences if there is overlap between neighboring n-mers based upon physical location. The blocks of unique and repeat n-mers are outputted in GFF format (http://www.sanger.ac.uk/resources/software/gff/spec.html). The GFF format file containing the coordinates of repeated exact match

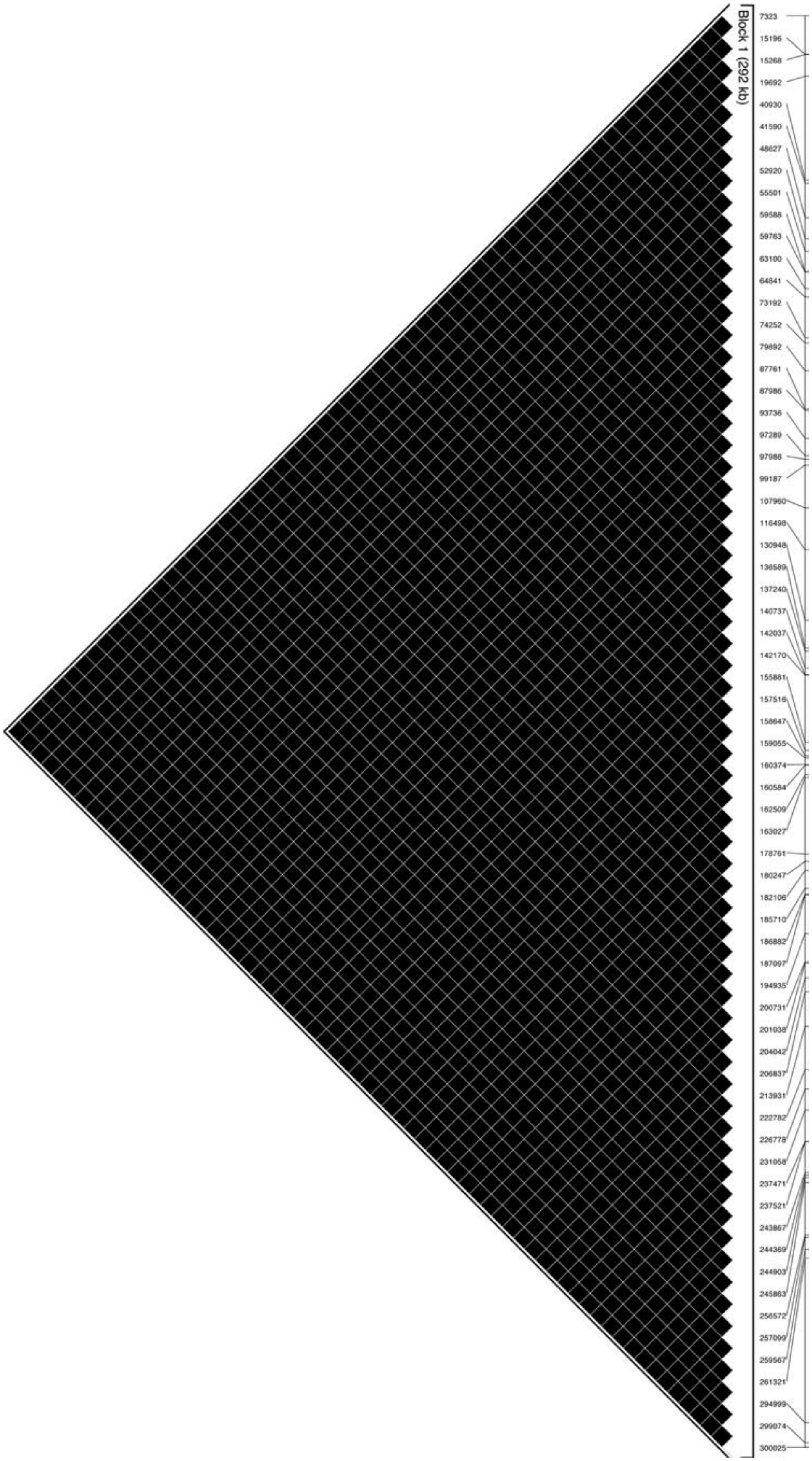

Figure 4 | **Haploview plot showing the results of the 4-gamete test.** The black blocks indicate fewer than 4 distinct 2-marker haplotypes for a pair of sites. We see no pairs of sites with 4 distinct 2-marker haplotypes. This result is consistent with a complete absence of historical recombination in the genomic region sequenced in the 39 worldwide *B. anthracis* strains.

30-mer sequences was then used to filter the strain sequences using a custom Perl script.

The second screening method consisted of masking those sequenced bases called in less than 80% of the sequenced samples. The final sequence files are contained within a .zip archive in Supplemental File 1. Supplemental Table 2 reports the genome coordinates and percent bases called for each sample sequenced. Supplemental Table 3 contains the position and genotype of all single nucleotide variants (SNVs) discovered in this study.

Phylogenetic Analyses. The PHYLIP package (v3.69) was used for all phylogenetic analyses[52]. UniqueMER filtered RA genome sequences for each strain were concatenated to create a single strain sequence in FASTA format. RA sequences were converted to PHYLIP format using Clustal X for subsequent analyses[53]. A custom Perl script (Phylip_neighbor_distance.pl) that called the PHYLIP program's dnadist and neighbor modules was used to generate a distance matrix and determine a neighbor-joining (NJ) tree for the RA datasets. A separate Perl script (Phylip_boot_distance.pl) that called the PHYLIP program's seqboot, dnadist, neighbor, and consense was used to generate 1000 replicate data sets for bootstrap analysis of the NJ trees. The PHYLIP program drawgram was used to draw the NJ trees. The Phylip program's dnapars and proml were use to confirm distance trees using parsimony and likelihood, respectively.

Population Genetic Analyses. All population genetic analyses were calculated using the popgen_fasta2.0.c code (Cutler DJ, unpublished work) on the 39 *B. anthracis* fasta files as previously described[22]. This code calculated the average number of pairwise differences and Watterson's estimator of the population mutation rate (Θ_w per site) for the entire sequenced region and different annotated SNV functional classes while accounting for missing data. A point estimate for Tajima's D was determined for all the data and different SNV functional classes. The statistical significance of these point estimates was determined relative to the standard neutral theory expectation, mainly a constant-sized population and mutation-drift equilibrium. Our linkage disequilibrium analysis of common single nucleotide variants (SNVs) included sites at greater than 10% frequency with genotype calls in at least 80% of samples analyzed. In order to analyze the list of common SNPs with McVean's LDHat program, a unique conversion script was written to generate the necessary *sites* and *locs* files. These files provide the input for **convert**. Within **convert**, all common SNVs as defined previously were analyzed. The output files from **convert,** in addition to a uniquely generated likelihood file, are then used as input for **interval**. **Interval** generates rates.txt and bounds.txt using an assigned start value for $2N_er$ that dictates the starting point for the RJMCMC. Multiple values for this starting parameter were tested, all of which provided identical output. Finally, **stat** was run to generate summary files for both the rates and bounds output files that were generated by **interval**. These summary files provide the mean, median, 2.5th percentile, and 97.5th percentile estimates of the recombination rates between each pair of SNPs, as well as the estimated locations of recombination rate changes in the region being analyzed. Haploview 4.2 was used to assess and visualize the linkage disequilibrium in the samples by performing the 4-gamete test[54]. The default 4-gamete color scheme was used, with black blocks representing less than 4 distinct 2-marker haplotypes.

1. Maiden, M. C. J. Multilocus sequence typing of bacteria. *Annual Review of Microbiology* **60**, 561–588 (2006).
2. Achtman, M. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annu Rev Microbiol* **62**, 53–70 (2008).
3. Didelot, X. & Maiden, M. C. J. Impact of recombination on bacterial evolution. *Trends in microbiology* **18**, 315–322 (2010).
4. Jolley, K. A., Wilson, D. J., Kriz, P., McVean, G. & Maiden, M. C. The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in Neisseria meningitidis. *Mol Biol Evol* **22**, 562–569 (2005).
5. Wirth, T. *et al.* The rise and spread of a new pathogen: seroresistant Moraxella catarrhalis. *Genome Res* **17**, 1647–1656 (2007).
6. Tanabe, Y., Sano, T., Kasai, F. & Watanabe, M. M. Recombination, cryptic clades and neutral molecular divergence of the microcystin synthetase (mcy) genes of toxic cyanobacterium Microcystis aeruginosa. *BMC Evol Biol* **9**, 115 (2009).
7. Touchon, M. *et al.* Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet* **5**, e1000344 (2009).
8. Didelot, X. & Falush, D. Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**, 1251–1266 (2007).
9. Didelot, X., Lawson, D., Darling, A. & Falush, D. Inference of Homologous Recombination in Bacteria Using Whole-Genome Sequences. *Genetics* **186**, 1435–1449 (2010).
10. Jackson, P. J. *et al.* Characterization of the variable-number tandem repeats in vrrA from different Bacillus anthracis isolates. *Appl Environ Microbiol* **63**, 1400–1405 (1997).
11. Keim, P. *et al.* Molecular diversity in Bacillus anthracis. *J Appl Microbiol* **87**, 215–217 (1999).
12. Smith, K. L. *et al.* Meso-scale ecology of anthrax in southern Africa: a pilot study of diversity and clustering. *J Appl Microbiol* **87**, 204–207 (1999).
13. Keim, P. *et al.* Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within Bacillus anthracis. *J Bacteriol* **182**, 2928–2936 (2000).
14. Smith, K. L. *et al.* Bacillus anthracis diversity in Kruger National Park. *J Clin Microbiol* **38**, 3780–3784 (2000).
15. Fouet, A. *et al.* Diversity among French Bacillus anthracis isolates. *J Clin Microbiol* **40**, 4732–4734 (2002).
16. Fasanella, A. *et al.* Molecular diversity of Bacillus anthracis in Italy. *J Clin Microbiol* **43**, 3398–3401 (2005).
17. Jackson, P. J., Hill, K. K., Laker, M. T., Ticknor, L. O. & Keim, P. Genetic comparison of Bacillus anthracis and its close relatives using amplified fragment length polymorphism and polymerase chain reaction analysis. *J Appl Microbiol* **87**, 263–269 (1999).
18. Price, L. B., Hugh-Jones, M., Jackson, P. J. & Keim, P. Genetic diversity in the protective antigen gene of Bacillus anthracis. *J Bacteriol* **181**, 2358–2362 (1999).
19. Radnedge, L. *et al.* Genome differences that distinguish Bacillus anthracis from Bacillus cereus and Bacillus thuringiensis. *Appl Environ Microbiol* **69**, 2755–2764 (2003).
20. Ko, K. S. *et al.* Identification of Bacillus anthracis by rpoB sequence analysis and multiplex PCR. *J Clin Microbiol* **41**, 2908–2914 (2003).
21. Helgason, E., Tourasse, N. J., Meisal, R., Caugant, D. A. & Kolstø, A. B. Multilocus sequence typing scheme for bacteria of the Bacillus cereus group. *Appl Environ Microbiol* **70**, 191–201 (2004).
22. Zwick, M. E. *et al.* Microarray-based resequencing of multiple Bacillus anthracis isolates. *Genome Biol* **6**, R10 (2005).
23. Read, T. D. *et al.* Comparative genome sequencing for discovery of novel polymorphisms in Bacillus anthracis. *Science* **296**, 2028–2033 (2002).
24. Read, T. D. *et al.* The genome sequence of Bacillus anthracis Ames and comparison to closely related bacteria. *Nature* **423**, 81–86 (2003).
25. Van Ert, M. N. *et al.* Global genetic population structure of Bacillus anthracis. *PLoS ONE* **2**, e461 (2007).
26. Cutler, D. J. *et al.* High-throughput variation detection and genotyping using microarrays. *Genome Research* **11**, 1913–1925 (2001).
27. Zwick, M. E., Kiley, M. P., Stewart, A. C., Mateczun, A. & Read, T. D. Genotyping of Bacillus cereus strains by microarray-based resequencing. *PLoS ONE* **3**, e2513 (2008).
28. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
29. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
30. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
31. Kimura, M. *The neutral theory of molecular evolution* (Cambridge University Press, Cambridge [Cambridgeshire] ; New York, 1983).
32. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
33. McVean, G. A. T. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science (New York, NY)* **304**, 581–584 (2004).
34. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science (New York, NY)* **310**, 321–324 (2005).
35. Watterson, G. A. The homozygosity test of neutrality. *Genetics* **88**, 405 (1978).
36. Hudson, R. R. & Kaplan, N. L. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164 (1985).
37. Maynard Smith, J. & Smith, N. H. Detecting recombination from gene trees. *Molecular Biology and Evolution* **15**, 590–599 (1998).
38. Chen, P. E. *et al.* Rapid identification of genetic modifications in *Bacillus anthracis* using whole genome draft sequences generated by 454 pyrosequencing. *PLoS ONE* **5** (2010).
39. Suerbaum, S. *et al.* Free recombination within Helicobacter pylori. *Proc Natl Acad Sci USA* **95**, 12619–12624 (1998).
40. Gomes, J. P. *et al.* Evolution of Chlamydia trachomatis diversity occurs by widespread interstrain recombination involving hotspots. *Genome Res* **17**, 50–60 (2007).
41. Chen, P. E. *et al.* Genomic characterization of the Yersinia genus. *Genome Biol* **11**, R1 (2010).
42. Holt, K. E. *et al.* High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. *Nature Genetics* **40**, 987–993 (2008).
43. Janes, B. & Stibitz, S. Routine markerless gene replacement in Bacillus anthracis. *Infection and Immunity* **74**, 1949 (2006).
44. Mironczuk, A. M., Kovacs, A. T. & Kuipers, O. P. Induction of natural competence in Bacillus cereus ATCC14579. *Microb Biotechnol* **1**, 226–235 (2008).
45. Kovacs, A. T., Smits, W. K., Mironczuk, A. M. & Kuipers, O. P. Ubiquitous late competence genes in Bacillus species indicate the presence of functional DNA uptake machineries. *Environ Microbiol* **11**, 1911–1922 (2009).
46. Hershberg, R. & Petrov, D. Evidence That Mutation Is Universally Biased towards AT in Bacteria. *PLoS Genet* **6**, e1001115 (2010).
47. Rocha, E. P. *et al.* Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol* **239**, 226–235 (2006).
48. Sawyer, S. A. & Hartl, D. L. Population genetics of polymorphism and divergence. *Genetics* **132**, 1161–1176 (1992).
49. Akashi, H. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. *Genetics* **139**, 1067–1076 (1995).
50. Kryazhimskiy, S. & Plotkin, J. The Population Genetics of dN/dS. *PLoS Genet* **4**, e1000304 (2008).

51. Priest, F. G., Barker, M., Baillie, L. W., Holmes, E. C. & Maiden, M. C. Population structure and evolution of the Bacillus cereus group. *J Bacteriol* **186**, 7959–7970 (2004).
52. Felstein, J. PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.* (2010).
53. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**, 4876–4882 (1997).
54. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).

Acknowledgements

The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the U.S. Government. The BDRD, NMRC authors are employees of the U.S. Government. This work was prepared as part of their official duties. Title 17 U.S.C. §105 provides that "Copyright protection under this title is not available for any work of the United States Government." Title 17 U.S.C. §101 defines a U.S. Government work as a work prepared by a military service member or employee of the U.S. Government as part of that person's official duties. This work was supported by the Transformational Medical Technologies Program under contract TMTI_IB06RSQ002 through the Defense Threat Reduction Agency to TDR and SS.

Author contributions

MZ, MKT, PC, HJ, and TR wrote the main manuscript text and MZ prepared figures 1–3. All authors reviewed the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: Drs. Zwick and Read are paid consultants with the Henry M. Jackson Foundation for Military Medicine with the Biological Defense Research Directorate, Naval Medical Research Center, Silver Spring, MD, USA.

License: This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/

How to cite this article: Zwick, M.E. *et al.* Genetic variation and linkage disequilibrium in *Bacillus anthracis*. *Sci. Rep.* **1**, 169; DOI:10.1038/srep00169 (2011).