Separating the effects of mutation and selection in producing DNA skew in bacterial chromosomes
Richard A Morton†1 and Brian R Morton*†2

Address: 1Department of Biology, McMaster University, 1280 Main Street West, Hamilton ON L8S 4K1, Canada and 2Department of Biological Sciences, Barnard College, Columbia University, 3009 Broadway, New York NY 10027, USA

Email: Richard A Morton - morton@mcmaster.ca; Brian R Morton* - bmorton@barnard.edu
* Corresponding author †Equal contributors

Abstract

Background: Many bacterial chromosomes display nucleotide asymmetry, or skew, between the leading and lagging strands of replication. Mutational differences between these strands result in an overall pattern of skew that is centered about the origin of replication. Such a pattern could also arise from selection coupled with a bias for genes coded on the leading strand. The relative contributions of selection and mutation in producing compositional skew are largely unknown.

Results: We describe a model to quantify the contribution of mutational differences between the leading and lagging strands in producing replication-induced skew. When the origin and terminus of replication are known, the model can be used to estimate the relative accumulation of G over C and of A over T on the leading strand due to replication effects in a chromosome with bidirectional replication arms. The model may also be implemented in a maximum likelihood framework to estimate the locations of origin and terminus. We find that our estimations for the origin and terminus agree very well with the location of genes that are thought to be associated with the replication origin. This indicates that our model provides an accurate, objective method of determining the replication arms and also provides support for the hypothesis that these genes represent an ancestral cluster of origin-associated genes.

Conclusion: The model has several advantages over other methods of analyzing genome skew. First, it quantifies the role of mutation in generating skew so that its effect on composition, for example codon bias, can be assessed. Second, it provides an objective method for locating origin and terminus, one that is based on chromosome-wide accumulation of leading vs lagging strand nucleotide differences. Finally, the model has the potential to be utilized in a maximum likelihood framework in order to analyze the effect of chromosome rearrangements on nucleotide composition.

Background

With the recent accumulation of complete bacterial genome sequences there has been increased attention to prokaryote chromosome organization. One prominent aspect of most of these genomes is that several features, such as nucleotide composition and coding strand bias, display an organization that is centered on the origin of replication [1]. In these chromosomes, as exemplified by Escherichia coli [2,3], replication initiates at a single origin (Ori) and proceeds bi-directionally to a terminus (Ter)
where the two forks meet [4]. This divides the chromosome into a replichore [5], defined as a chromosome with two oppositely replicated halves (or replication arms), within each of which there is a leading and lagging strand such that one DNA strand is leading within one replication arm but lagging within the other. Many bacterial genomes display a compositional asymmetry between the two DNA strands within a replication arm meaning that Parity Rule 2, which stipulates that the frequencies of A and T are equal as are the frequencies of G and C, is violated [6-8]. Observed strand asymmetry, or skew, in base composition is either a purine (G and A) or a keto (G and T) excess on one strand and the leading strand in one replication arm shows the same skew as the leading strand in the other despite the fact that they are opposite genome strands.

One issue that has arisen from these observations is the cause of the compositional asymmetry between strands, with evidence having been presented for contributions from both mutation and selection [1,9]. Many of the prokaryotes that have a replichore structure also have a bias towards coding genes, particularly 'essential' genes [1], on the leading strand [6,8,10-12] suggesting that composition asymmetry could result from selection and/or transcription-coupled mutation and repair processes [9,10,13]. There is also evidence, though, that leading and lagging strands differ in mutation bias [7,8,14], which has interesting and important implications for genome evolution [14,15]. Although there have been estimates of the contribution of mutation to skew in several genomes that suggest a role for selection [7] the issue has not been studied within a statistical framework.

Replichores with strand asymmetry have also been exploited to make inferences about the location of an origin of replication when the origin has not been mapped experimentally, which is the case in the vast majority of sequenced genomes. The compositional difference between leading and lagging strands, and the replichore structure in general, means that the two DNA strands have complementary composition biases in the two replication arms of these genomes. Plotting composition skew along a sliding window leads to a characteristic pattern in which the origin lies at a point where a given measurement of strand asymmetry switches between positive and negative values. This type of graphical approach has been used frequently to infer the location of replication origins [4,16-20] but these approaches have the disadvantage that determining the existence of skew and where it switches strand is subjective [9]. A more objective linear discriminant analysis has also been developed [21], but this method does not account for gene density nor does it utilize intergenic regions [9].

We develop a simple bipartition model (Methods) that exploits the existence of a replichore structure with strand asymmetry and provides for analyses within a maximum likelihood (ML) framework. In the current study we will apply the model to bacterial chromosomes (see Table 1) in two analyses. The first is the identification of two peaks on a likelihood surface, the locations of the putative origin and terminus of replication. Additional information, such as the organization of ribosomal RNA (rRNA) genes [22], can then be used to infer which of the two locations is the origin. In this study we assign the origin to maximize the arrangement of rRNA genes on the resulting leading strands [22] and then compare the location of our putative origin to "origin-associated" genes that are located at or close to the replication origin in certain genomes [23]. The second application is a statistical analysis of the contribution of mutation to nucleotide composition skew across bacterial species.

**Results and discussion**

We applied the bipartition model to 352 fully sequenced bacterial chromosomes. An assessment of the mutational (which we call R-dependent) component of compositional skew (see Methods) requires an identification of the origin and terminus of replication. Since these have not been empirically identified in most genomes we first use the model to generate a maximum likelihood estimation of the two loci and discuss the accuracy of this approach. Once the putative origin and terminus have been identified for each chromosome we use the model to quantify the degree to which the mutational difference between leading and lagging strands generates a skew.

**Table 1: Chromosomes analyzed from eubacterial phyla**

| Phylum                  | Class | Number |
|-------------------------|-------|--------|
| Actinobacteria          |       | 23     |
| Bacteroidetes           |       | 5      |
| Chlamydia               |       | 11     |
| Cyanobacteria           |       | 17     |
| Deinococcus-Thermus     |       | 4      |
| Firmicutes              |       | 79     |
|                         |       |        |
|                         | Mollicutes | 16     |
|                         | Lactobacillales | 27     |
|                         | Clostridia  | 7      |
|                         | Bacillales  | 29     |
| Proteobacteria          |       | 197    |
|                         | Alphaproteobacteria | 55     |
|                         | Betaproteobacteria | 38     |
|                         | Deltaproteobacteria | 11     |
|                         | Epsilonproteobacteria | 8     |
|                         | Gammaproteobacteria | 85     |
| Spirochaetes            |       | 6      |
| Others                  |       | 10     |

1The two phyla with the largest sample sizes are shown sub-divided by Class representation.
Replication arms comprise half of most bacterial chromosomes

We expect that if the unordered pair of sites identified by our model as the potential origin and terminus ([S₁, S₂]ₘₐₓ, see Methods) is an accurate estimation of the replication origin and terminus then these two sites should divide the chromosome roughly in half. (This chromosome division does not depend upon the assignment of Sᵢₘₐₓ as an origin or terminus, which will be the subject of the next section.) The existence of such a physical balance is the basis of the Adopt-Adapt model [24,25,4], which proposes that this balance would guarantee a synchronous completion of replication by the two forks. The distribution of chromosome divisions, measured by the statistic C_d (see Equation 10, Methods) is shown in Figure 1 for the 326 circular chromosomes in which the 95 percent confidence interval of C_d was less than 20% of the chromosome. Linear chromosomes were also excluded since the method of calculating the confidence interval cannot be applied to them. We took C_d between 0 and 0.2 as indicating a relatively equal division of chromosome arms and > 0.2 as an arbitrary measure of an inequitable division: these are chromosomes in which [S₁, S₂]ₘₐₓ generated a replication arm less than 40% of the total length. Of the 326 chromosomes analyzed, 31 (9.5%) had an inequitable division by our definition (see Table 2). Overall, this result agrees with the prediction of the ‘Adopt-Adapt’ model. The advantage to such an arrangement would presumably be that it results in the shortest possible replication time, which would otherwise be limited by the time taken to replicate the longer of the two arms.

Given an expectation for physical balance of chromosomes, it is possible that the 31 chromosomes with an inequitable division have undergone recent rearrangements. Large indels in either replication arm or inversions that include either the origin or the terminus could lead to deviations from equitable distribution. (Inversions that do not include the real replication origin could influence the ability to detect significant compositional skew but will not affect C_d.) This cannot be investigated for all 31 cases since many chromosomes have not been studied in great detail or do not have close relatives for comparison, but evidence suggests that at least some of these chromosomes have undergone the predicted rearrangements. For example, *Pseudomonas aeruginosa* PA01 has a large inversion encompassing about 1/3 the genome [26], *Yersinia pestis* str. 91001 has significant inversions surrounding the putative origin [27] and *Prochlorococcus marinus* MIT9313 has several genome rearrangements relative to strain MED4 and is also much larger [28]. *Candidatus Blochmannia pennsylvanicus* str. BPEN has undergone significant gene loss relative to *B. floridans* [29], *Halobacterium sp. NRC-1* contains 91 Insertion Sequences from 12 different families [30] and *Candidatus Pelagibacter ubique* HTCC1062 has a very small genome that appears to have undergone numerous recent deletions [31]. These rearrangements suggest the possibility that the origin and terminus in these chromosomes do, in fact, yield a noticeably inequitable division. Although it seems likely that selection would favor an equitably divided genome, it is not known how rapidly a chromosome would regain equitable division following such a rearrangement and the genomes in Table 2 may be evolving towards a more equitable division. Alternatively, it could be that the genome assembly was performed incorrectly (see [32] for such an example) or that in these cases skew is not at equilibrium as a result of genome rearrangement and [S₁, S₂]ₘₐₓ does not represent the origin and terminus. If this is the case then any method that utilizes composition skews to estimate the origin of replication, whether our ML approach or a graphical approach, would be misled.

Locating the putative replication origins of bacterial chromosomes

We used [S₁, S₂]ₘₐₓ to assign a putative origin/terminus pair (which we then call [P₀, P₁]ₘₐₓ where P₀ₘₐₓ is either S₁ₘₐₓ or S₂ₘₐₓ and P₁ₘₐₓ is the other location) for the 352 chromosomes in our dataset. As described in the Methods we used several different methods to accomplish this;
each are described separately. Where possible we can compare our results to empirically identified replication origins to assess the accuracy of our approach.

**Linear chromosomes**

The five linear chromosomes are relatively simple since an origin has been annotated for each at the center of the genome and these annotations appear to be reliable: the annotations for *Borrelia burgdorferi* and *B. garnii* are based on empirical evidence [17] as are the annotations for the two Streptomyces taxa [33], and in the case of *A. tumefaciens* the annotation is based on an analysis of the repABC genes [34]. Therefore, we can assign POML and PTML for the linear chromosomes in the dataset using the annotated origins and assess our method. We find a putative origin near the annotated origin in each of the five linear chromosomes (Table 3); in three it is within 0.1% of the genome length of the annotated origin while in the other two it is within 3% of the genome length. In each case the other site is near the end of the chromosome. The data indicate that the bipartition model locates the origin and terminus accurately in these five chromosomes.

**Ribosomal RNA genes in circular chromosomes**

There is strong evidence that circular chromosomes are organized such that the ribosomal RNA (rRNA) genes tend to be located on the leading strand of replication regardless of where they are distributed along the length of the chromosome [22]. We used this as a basis for assigning [P0, P1][ML] in the 319 primary and 28 secondary circular chromosomes in our dataset. If S1[ML] is assigned as the origin then in the chromosome portion S1[ML] → S2[ML] the + strand is the leading strand while in the S2[ML] → S1[ML] portion the – strand is leading. Assigning S2[ML] as the origin reverses this leading strand assignment. We calculated the proportion of rRNA coded on the leading strand in the two possible organizations and assigned the origin based on which of the two resulted in a majority of rRNA genes

---

**Table 2: Chromosomes with unequal replication arms**

| Taxon2 | Accession | Phylum; Class3 |
|--------|-----------|----------------|
| Acidobacteria bacterium Ellin345 | NC_008009 | Acidobacteria |
| Aquifex aeolicus VF5 | NC_000918 | Aquificae |
| Aster yellows witches’-broom phytoplasma AYYWB | NC_007716 | Firmicutes; Mollicutes |
| Bordetella parapertussis | NC_002928 | Proteobacteria; Betaproteobacteria |
| Bordetella pertussis Tohama I | NC_002929 | Proteobacteria; Betaproteobacteria |
| Burkholderia cenocepacia AU1054 | NC_008060 | Proteobacteria; Betaproteobacteria |
| Burkholderia thailandensis E264 | NC_007651 | Proteobacteria; Betaproteobacteria |
| Candidatus Blochmannia pennsylvanicus str. BPEN | NC_007292 | Proteobacteria; Gammaproteobacteria |
| Candidatus Pelagibacter ubique HTCC1062 | NC_007205 | Proteobacteria; Alphaproteobacteria |
| Desulfitobacterium hafniense YSI | NC_007907 | Firmicutes; Clostridia |
| Francisella tularensis subsp. holarctica | NC_007880 | Proteobacteria; Gammaproteobacteria |
| Haemophilus ducreyi 35000HP | NC_002940 | Proteobacteria; Gammaproteobacteria |
| Heliothrix hepaticus ATCC S1499 | NC_004917 | Proteobacteria; Epsilonproteobacteria |
| Idiomarina loihiensis L2TR | NC_006512 | Firmicutes; Mollicutes |
| Mycoplasma hyopneumoniae J | NC_007295 | Firmicutes; Mollicutes |
| Mycoplasma mobile 163K | NC_006908 | Firmicutes; Mollicutes |
| Mycoplasmas penetrans HF2 | NC_004432 | Firmicutes; Mollicutes |
| Nitrrosomonas europaea ATCC 19718 | NC_004757 | Proteobacteria; Betaproteobacteria |
| Prochlorococcus marinus MIT9313 | NC_005071 | Cyanobacteria |
| Pseudomonas aeruginosa PAO1 | NC_002516 | Proteobacteria; Gammaproteobacteria |
| Pseudomonas putida KT2440 | NC_002947 | Proteobacteria; Gammaproteobacteria |
| Shigella dysenteriae Sd197 | NC_007606 | Proteobacteria; Gammaproteobacteria |
| Silicibacter pomeroyi Megaplasmid | NC_006569 | Proteobacteria; Alphaproteobacteria |
| Sinorhizobium meliloti pSymA | NC_003037 | Proteobacteria; Alphaproteobacteria |
| Sodalis glossinidius str. ‘morsitans’ | NC_007712 | Proteobacteria; Gammaproteobacteria |
| Synecochococcus elongatus PCC7942 | NC_007604 | Cyanobacteria |
| Thermus thermophilus HB8 | NC_006461 | Deinococcus-Thermus |
| Wolbachia endosymbiont of D. melanogaster | NC_002978 | Proteobacteria; Alphaproteobacteria |
| Xylella fastidiosa 9a5c | NC_002488 | Proteobacteria; Gammaproteobacteria |
| Yersinia pestis Antiqua | NC_008150 | Proteobacteria; Gammaproteobacteria |
| Yersinia pestis MED | NC_005810 | Proteobacteria; Gammaproteobacteria |

1One replication arm less than 40% of the total chromosome length (see text).

2Primary chromosome unless otherwise indicated.

3Class is given only for Firmicutes and Proteobacteria as in Table 1.
on the leading strand. In 14 chromosomes there was no annotation of rRNA genes while in 8 others we could not assign an origin because both possibilities yielded 50% rRNA genes on each strand. For the remaining 325 chromosomes, 304 had 100% of the rRNA genes on the putative leading strand, 18 others had > 75% of the rRNA genes on the putative leading strand and the 3 others had 60% of the rRNA genes on the putative leading strand. An interesting point involves assignment of the origin to the middle of the linear Borrelia chromosomes. In Borrelia garinii only 20% of the rRNA genes are on the leading strand. The 23S rRNA genes in Borrelia garinii are inverted with respect to Borrelia burgdorferi where 97% of the rRNA sites are on the leading strand.

Evidence other than RNA genes
For the 22 circular chromosomes noted above for which an origin could not be assigned based on the distribution of rRNA genes, we designated [P₀, P₁]ML based on assigning the leading strand in each replication arm as the strand for which the replication-induced effect was inferred to make C > C at four-fold degenerate (D₄) sites. The rationale for this is discussed in the Methods and necessarily limits the conclusions we can draw about composition bias in these chromosomes.

Identification of genes consistently located near the origin of bacterial chromosomes
One interesting feature of the circular chromosomes in our dataset is the existence of genes that may function in segregating the two replication products during cell division [23] and which appear to be located near the origin of replication in many chromosomes. In this section we compare the locations of P₀, P₁]ML in each chromosome to these genes in order to determine the degree of concurrence. We will do this separately for the 28 secondary and 319 primary circular chromosomes; the 5 linear chromosomes will not be analyzed in this section. This comparison is not essential for our identification of [P₀, P₁]ML but it provides a way of both assessing our model as well as the proposal that these genes do tend to be located near the replication origin.

Secondary chromosomes
In the secondary chromosomes, the segregation system involves either the parA and parB genes, along with a co-localized cis element (parS), which is bound by the ParB protein, or the repA, repB (which are apparently parA and parB homologs respectively, see [35]) and repC genes. We will refer to this generally as par/rep system; it is found in all of the secondary chromosomes in this study. The location of the par/rep genes in each of the 28 secondary chromosomes was determined by annotation if possible, while for the chromosomes in which neither was annotated we used a BLAST search to locate the gene. In 26 of the 28 secondary chromosomes we find that P₀, P₁]ML is within 1% of the genome length of par/rep and in one of the other two chromosomes the distances is 2.6% of the chromosome length (Table 4). In the remaining case, Rhodobacter sphaeroides chromosome 2, it is P₀, P₁]ML and not P₀, P₁]ML that is near par/rep, although the assignment of P₀ was based on it resulting in all of the rRNA genes being on the leading strand (see Additional file 1). Therefore, either P₀, P₁]ML is near the par/rep location and none of the rRNA genes are on the leading strand or P₀, P₁]ML is near the par/rep location and all of the rRNA genes are on the leading strand. It is possible that a recent inversion or other chromosome rearrangement resulted in a separation of the segregation genes and the origin, leading to an incorrect assignment of the origin based on rRNA distribution. Overall, the co-localization of P₀, P₁]ML and the par/rep genes in all but this one case provides strong evidence that the bipartition model is accurately locating the replication origin.

Primary chromosomes
For the analysis of the 319 primary, circular chromosomes we excluded 21 chromosomes for which the statistical significance of [S₁, S₂]ML is uncertain; 14 of them because there was no evidence for a replication-induced effect on composition, and the other 7 because the confidence limit of the genome division parameter Cd spanned more than 20% of the chromosome. For the remaining 298 chromosomes we used what we will call an "origin gene" approach to assess the location of P₀, P₁]ML, an approach that

Table 3: Comparison of the origin of replication and the putative ML origin for each of the five linear chromosomes

| Bacterial Species | Accession | Origin Signal | Origin Location | [S₁, S₂]ML | P₀, P₁]ML | Distance |
|-------------------|-----------|---------------|----------------|------------|-----------|----------|
| Streptomyces avermitilis | NC_003155 | oriC | 0.586 | 0.569, 0.994 | 0.569 | 0.017 |
| Streptomyces coelicolor | NC_003888 | oriC | 0.493 | 0.463, 0.987 | 0.463 | 0.030 |
| Borrelia burgdorferi | NC_001318 | dnaA-N | 0.503 | 0.000, 0.503 | 0.503 | 0.000 |
| Borrelia garinii | NC_006156 | dnaA-N | 0.509 | 0.000, 0.510 | 0.510 | 0.001 |
| A. tumefaciens² | NC_003305 | repA | 0.494 | 0.000, 0.493 | 0.493 | 0.001 |

¹Location of the annotated origin as a fraction of chromosome length starting at the NCBI site 1.
²Location of [S₁, S₂]ML as a fraction of chromosome length.
³P₀, P₁]ML as inferred by which of S₁, S₂]ML is closest to the annotated origin.
⁴Distance between the annotated origin and P₀, P₁]ML.

[Page 5 of 17 (page number not for citation purposes)]
is similar to what was applied to the secondary chromosomes. Unlike the secondary chromosomes, there is no single cluster of genes that can be used to locate the origin in all of the primary genomes in the dataset. Despite this lack of a universal gene cluster, evidence indicates that the origin of replication in these primary chromosomes is frequently located nearby any or all of several genes which may actually have formed an ancestral origin gene “cluster” [36]. These genes, which we will refer to as “origin genes”, are parA, parB, gidA, gidB, yidC, yidD, rpmA, rpmH, dnaA and dnaN [23]. These genes vary in location with respect to one another and thus cannot always be used to determine a single chromosome location, but despite this, evidence from chromosomes where these genes have been studied indicates that at least one of these genes is close to the origin in any given chromosome [23].

We used these so-called origin genes to assess \( P_{O}^{ML} \) as follows. NCBI annotations were used to locate all genes annotated with any of the 10 gene names. Since these genes are not located together across all chromosomes we treated them as 5 pairs that tend to be co-located: parA and parB, gidA and gidB, yidC and yidD, rpmA and rpmH, and dnaA and dnaN. A chromosome was scored as having a specific gene pair if both genes in that pair were annotated and the mid-points were within 5 kb of each other or if only one gene of the pair was annotated (under the assumption that the second gene may not have been annotated). We then scored \( P_{O}^{ML} \) as being located near that specific pair if it was within 1% of the chromosome length of the midpoint between the two genes or from the midpoint of the single annotated gene. Of the 298 chromosomes, 57 had none of the 10 genes annotated and thus could not be assessed. For the remaining 241, Table 5 shows the number of gene pairs that were scored for each chromosome along with how many chromosomes had \( P_{O}^{ML} \) near any pair. Overall 154 chromosomes (63.9%) have the putative origin within 1% of the genome length of one of the pairs and of the 156 chromosomes that have 3 or more gene pairs, 115 (73.7%) have the putative origin within 1% of the length from a gene pair. Of the 9 chromosomes with all 5 gene pairs located, 8 have \( P_{O}^{ML} \) near at least one pair and in the one exception (Azoarcus sp. EbN1), all 5 gene pairs were close to one another and the putative origin was between 5.1% and 6.0% of the pairs. In addition, in cases where there was

| Bacterial Species | Accession | \( \text{parA} / \text{repA}\) | \( P_{O}^{ML} \) | Distance |
|-------------------|-----------|----------------|--------------|-----------|
| Brucella abortus   | NC_006933 | 0.998          | 0.999        | 0.001     |
| Brucella melitensis, 16M | NC_003318 | 0.079          | 0.079        | 0.000     |
| Brucella melitensis, Abortus 2308 | NC_007624 | 0.998          | 0.999        | 0.001     |
| Brucella suis      | NC_004311 | 0.998          | 0.999        | 0.001     |
| Burkholderia sp. 383, Chr 2    | NC_007511 | 0.002          | 0.000        | 0.002     |
| Burkholderia sp. 383, Chr 3    | NC_007509 | 0.996          | 0.999        | 0.005     |
| Burkholderia cenocepacia, Chr 2 | NC_008061 | 0.879          | 0.878        | 0.001     |
| Burkholderia cenocepacia, Chr 3 | NC_008062 | 0.315          | 0.320        | 0.005     |
| Burkholderia mallei   | NC_006349 | 0.997          | 0.999        | 0.001     |
| Burkholderia pseudomallei, 1710b | NC_007435 | 0.576          | 0.578        | 0.002     |
| Burkholderia pseudomallei, K96243 | NC_006351 | 0.998          | 0.999        | 0.003     |
| Burkholderia thailandensis | NC_007650 | 0.998          | 0.999        | 0.002     |
| Burkholderia xenovorans, Chr 2 | NC_007952 | 0.000          | 0.006        | 0.006     |
| Burkholderia xenovorans, Chr 3 | NC_007953 | 0.981          | 0.979        | 0.002     |
| Photobacterium profundum | NC_006371 | 0.999          | 0.999        | 0.001     |
| Ralstonia eutropha     | NC_007348 | 0.775          | 0.778        | 0.003     |
| Ralstonia metallidurans | NC_007974 | 0.943          | 0.946        | 0.003     |
| Ralstonia solanacearum | NC_003296 | 0.001          | 0.002        | 0.001     |
| Rhodobacter sphaeroides | NC_007494 | 0.996          | 0.513        | 0.483     |
| Silicibacter sp. TM1040  | NC_008043 | 0.517          | 0.519        | 0.002     |
| Silicibacter pomeroyi   | NC_006569 | 0.678          | 0.668        | 0.010     |
| S. meliloti, pSymA     | NC_003037 | 0.999          | 0.991        | 0.008     |
| S. meliloti, pSymB     | NC_003078 | 0.034          | 0.060        | 0.026     |
| Vibrio cholerae       | NC_002506 | 0.999          | 0.999        | 0.005     |
| Vibrio fischeri       | NC_006841 | 0.999          | 0.999        | 0.003     |
| Vibrio parahaemolyticus | NC_004605 | 0.999          | 0.000        | 0.001     |
| Vibrio vulnificus CMCP6 | NC_004460 | 0.689          | 0.692        | 0.003     |
| Vibrio vulnificus, YO16 | NC_005140 | 0.999          | 0.002        | 0.002     |

\( \text{1parA or repA location (see text), as a fraction of chromosome length starting at the NCBI site.} \)
\( \text{2}P_{O}^{ML}\text{ as inferred from rRNA sites or } \sigma_{GC} > 0 \text{ (see text).} \)
\( \text{3Distance between parA/repA location and } P_{O}^{ML}. \)
more than one gene pair annotated, the most common category in Table 5 is that in which \( P_O \) is located within 1% of all identified gene pairs. This indicates that the origin genes were all clustered and very close to our putative origin. The results confirm that the replication origin is frequently near one or more of the set of "origin genes".

**R-dependent component of skew across bacterial chromosomes**

Once we have identified \([P_O, P_I]^{ML}\) for each chromosome we can use our model to calculate the effect of leading vs lagging strand mutation differences, the R-dependent component of skew. As discussed in the Methods, the R-dependent component is best estimated by using sites least affected by selection. The intergenic (IG) regions and four fold degenerate (D4) sites of CDS regions are possible choices. We used relative \( \sigma \) values (\( R_G \) and \( R_T \), see Equation 9, Methods) to compare the R-dependent component of skew across all 352 bacterial chromosomes. These measures represent the proportional increase of G and T, respectively, on the leading strand as a result of mutational differences between the leading and lagging strands.

There is a strong correlation between IG and D4 for both \( R_G \) (Slope = 0.68; \( R^2 = 0.76, P < 10^{-6} \)) and \( R_T \) (Slope = 0.46; \( R^2 = 0.73, P < 10^{-6} \)), indicating that they give similar estimates of the relative contribution of the R-dependent mutation effect to skew. It also suggests that transcription-coupled mutation effects are not strongly affecting the estimates of R-independent effects at D4 sites. However, the data also show that for most chromosomes the absolute value of the skew is stronger for D4 sites than IG sites, for both \( R_G \) and \( R_T \). We propose, based on the compact nature of intergenic sites in bacterial genomes and the existence of regulatory sequences within these regions, that this is most likely due to selective constraints on many intergenic sites similar to recent findings from Drosophila [37], and that the D4 sites provide the more accurate estimation of the contribution of the R-dependent component. Therefore, we will use these sites to examine this component of skew across bacterial chromosomes.

Given that there is also a strong correlation between D4 sites coded on the + and - strands across the 352 bacterial chromosomes (for \( R_G \), \( R^2 = 0.90 \); for \( R_T \), \( R^2 = 0.94 \)), we calculated the average of the two strands for both \( R_G \) and \( R_T \) to obtain a single estimate to assess the relative strength of replication-induced skew (Figure 2). We also calculated a measure of overall skew \( \sqrt{R_G^2+R_T^2} \) and those chromosomes with the strongest and weakest overall skew are shown in Table 6. There are 208 chromosomes that have a significantly stronger G/C than T/A skew (below the \( R_G = R_T \) line) and 90 that have a significantly stronger T/A than G/C skew (above the \( R_G = R_T \) line). \( R_G \) and \( R_T \) are not significantly different in the remaining 54. This suggests that mutation bias tends to generate a stronger G/C skew across microbial chromosomes [7] but a statistical analysis is confounded by the fact that the points are not independent due to shared ancestry as well as the fact that the species that have been sequenced have not been sampled randomly.

![Strand asymmetry across bacterial chromosomes](image)

**Table 5: Number of chromosomes in which the putative origin is located near one or more origin gene pairs**

| Gene Pairs | 0 | 1 | 2 | 3 | 4 | 5 | > 5 | Total |
|------------|---|---|---|---|---|---|-----|-------|
| 0          | 57 | - | - | - | - | - | 57  |
| 1          | 24 | 7 | - | - | - | 7 | 31  |
| 2          | 22 | 11| 21| - | - | 32| 54  |
| 3          | 17 | 3 | 23| 24| - | 50| 67  |
| 4          | 23 | 11| 12| 3 | 31| - | 57  |
| 5          | 1  | 2 | 0 | 1 | 0 | 5 | 8   |

1 Number of "origin gene" pairs found in any given chromosome, as defined in the text.
2 Number of chromosomes for which the specified number of origin gene pairs is found within 1% of the chromosome length from \( P_O^{ML} \). When multiple origin gene pairs are found near \( P_O^{ML} \) it indicates that the gene pairs are close together on the chromosome.
3 Total number of chromosomes with that number of annotated gene pairs regardless of whether or not any pair is close to \( P_O^{ML} \).
The G/C skew is almost exclusively biased towards G, which agrees with the observation of Lobry and Sueoka [7] who found that G > C on the leading strand in almost all of the bacterial genomes they surveyed, the one exception being the linear chromosome of *Streptomyces coelicolor*. Since we used the criterion of G > C on the leading strand to assign POML for 22 of the circular chromosomes, we excluded these from an assessment of the pattern of G/C bias. Of the remaining 330 chromosomes, only 9 show a significant leading strand skew towards C. These are *Bifidobacterium longum*, *Thermobifida fusca* YX, *Deinococcus radiodurans* R1, *Tropheryma whipplei* TW08/27, *T. whipplei* str. Twist, *Mycoplasma mobile* 163K and *M. penetrans* HF-2 as well as the two linear *Streptomyces* chromosomes, *S. avermitilis* MA-4680 and *S. coelicolor*. We also find that among those chromosomes with a significant T/A skew (above or below the RT = 0 line), 245 have a significant skew towards T and 65 a significant skew towards A. Overall, there were 237 chromosomes with a significant G&T bias, 60 with a G&A bias, 5 with a C&T bias and 4 with a C&A bias on the leading strand. This analysis supports the finding of Lobry and Sueoka [7] that the G&T

| Species                                                   | Group                  | R_G | R_T   | Skew^1 |
|-----------------------------------------------------------|------------------------|-----|-------|--------|
| *Candidatus Blochmannia floridanus*                       | Gammaproteobacteria    | 0.433 | 0.107 | 0.446  |
| *Clostridium acetobutylicum* ATCC 824                     | Firmicutes             | 0.426 | -0.087| 0.435  |
| *Ehrlichia canis* str. Jake                              | Alphaproteobacteria.   | 0.346 | 0.226 | 0.413  |
| *Ehrlichia ruminantium* str. Welgevonden                 | Alphaproteobacteria     | 0.360 | 0.191 | 0.408  |
| *Ehrlichia ruminantium* str. Arkansas                    | Alphaproteobacteria     | 0.360 | 0.191 | 0.407  |
| *Ehrlichia ruminantium* str. Gardel                      | Alphaproteobacteria     | 0.353 | 0.190 | 0.401  |
| *Ehrlichia ruminantium* str. Welgevonden                 | Alphaproteobacteria     | 0.349 | 0.188 | 0.397  |
| *Lactobacillus salivarius* subsp. salivarius* UCC118      | Firmicutes             | 0.362 | -0.121| 0.382  |
| *Barrela burgdorferi* B3I                                 | Spirochaetes            | 0.229 | 0.300 | 0.378  |
| *Clostridium perfingens* str. 13                         | Firmicutes             | 0.350 | -0.086| 0.360  |
| *Barrela garinii* PBI                                     | Spirochaetes            | 0.207 | 0.287 | 0.354  |
| *Xylella fastidiosa* Temecula I                          | Gammaproteobacteria     | 0.254 | 0.228 | 0.341  |
| *Bartonella quintana* str. Toulouse                       | Alphaproteobacteria     | 0.300 | 0.080 | 0.311  |
| *Clostridium tetani* E88                                 | Firmicutes             | 0.273 | -0.137| 0.305  |
| *Bartonella henselae* str. Houston-I                     | Alphaproteobacteria     | 0.284 | 0.080 | 0.295  |
| *Xylella fastidiosa* 9a5c                                 | Gammaproteobacteria     | 0.197 | 0.218 | 0.294  |
| *Lactobacillus acidophilus* NCFM                         | Firmicutes             | 0.286 | -0.024| 0.287  |
| *Buchnera aphidicola* str. Bp                            | Gammaproteobacteria     | 0.254 | 0.124 | 0.282  |
| *Lactobacillus johnsonii* NCC 533                        | Firmicutes             | 0.260 | -0.042| 0.263  |
| *Carboxydothermus hydrogenoformans* Z-290I                | Firmicutes             | 0.260 | 0.000 | 0.260  |

Table 6: Chromosomes with extreme skew

| Species                                                   | Group                  | R_G | R_T   | Skew^1 |
|-----------------------------------------------------------|------------------------|-----|-------|--------|
| *Corynebacterium efficiens* YS-314                        | Actinobacteria         | 0.023 | 0.019 | 0.030  |
| *Nostoc sp. PCC 7120*                                     | Cyanobacteria          | 0.028 | 0.005 | 0.028  |
| *Sinorhizobium meliloti* 1021 plasmid pSymA               | Alphaproteobacteria     | 0.022 | 0.014 | 0.026  |
| *Mycobacterium avium* subsp. paratuberculosis* K-10      | Actinobacteria         | 0.002 | 0.025 | 0.025  |
| *Aeromonas* bacter dehalogenans 2CP-C                     | Deltaproteobacteria    | 0.022 | -0.009| 0.024  |
| *Mycoplasma hyopneumoniae* 232                           | Firmicutes             | 0.017 | 0.016 | 0.023  |
| *Mycoplasma synoviae* S3                                 | Firmicutes             | -0.012| 0.016 | 0.020  |
| *Wigglesworthia glossinidia*                             | Gammaproteobacteria    | 0.016 | 0.011 | 0.020  |
| *Mycoplasma mycoides* subsp. mycoides SC str. PG1        | Firmicutes             | 0.017 | 0.009 | 0.019  |
| *Mycoplasma gallisepticum* R                             | Firmicutes             | -0.014| 0.009 | 0.016  |
| *Aquifex aeolicus* VF5                                   | Aquificae              | -0.006| -0.014| 0.015  |
| *Mycoplasma hyopneumoniae* 7448                          | Firmicutes             | -0.012| -0.002| 0.013  |
| *Thermosynechococcus elongatus* BP-I                     | Cyanobacteria          | -0.005| -0.010| 0.012  |
| *Synechocystis sp. PCC 6803                              | Cyanobacteria          | 0.010 | -0.002| 0.010  |
| *Synechococcus sp. JA-2-38a(2–13)                        | Cyanobacteria          | 0.009 | -0.001| 0.009  |
| *Baumannia cicadellinicola* str. Hc                       | Gammaproteobacteria    | -0.003| -0.005| 0.006  |
| *Anabaena variabilis* ATCC 29413                         | Cyanobacteria          | -0.005| 0.003  | 0.006  |
| *Mycoplasma pneumoniae* M129                             | Firmicutes             | 0.003 | 0.003 | 0.004  |
| *Synechococcus sp. JA-3-3Ab                               | Cyanobacteria          | 0.001 | 0.004 | 0.004  |
| *Gloeobacter violaceus* PCC 7421                         | Cyanobacteria          | 0.003 | -0.002| 0.004  |

^1 Overall skew given by SQRT(R_G^2 + R_T^2)
(keto) skew is more common than G&A (purine) skew on the leading strand. Statistical analysis is again confounded by non-independence of points.

Despite the limitations that arise from the non-independence of the chromosomes one conclusion that can be drawn from the data is that the chromosomes with a purine (G&A) skew are predominantly Gram-Positive bacteria (figure 3). Lobry and Sueoka [7] previously noted the exception of the Gram-Positive Lactococcus lactis and Staphylococcus aureus to the general trend of keto skew. We show that it is a common feature across the Firmicutes, with the exception of the genus Mycoplasma with 12 representative chromosomes, all of which show very little overall R-dependent effect. Of the 64 genomes that have significant strand average skew towards A in Figure 2, 53 are Firmicutes and only 4 of the 79 sequenced Gram-Positive chromosomes show significant strand average skew towards T (Mycoplasma penetrans HF-2, Bacillus subtilis subsp. subtilis str. 168, Bacillus licheniformis ATCC 14580, and Thermoaerobacter tengcongensis MB4). These trends are difficult to assess without a general knowledge of how rapidly mutational biases change and the degree of relationship in the genomes being compared, but it does appear that the Gram-Positive chromosomes tend to show a different general pattern of skew than other prokaryotes.

The difference in composition bias between the Firmicutes and the other bacterial lineages may be related to a general difference in the mode of replication. It has been noted previously that those species that have a leading strand bias towards A over T have a polC homolog in addition to a dnaE homolog [20]. These authors point out that proofreading in species with only a dnaE homolog involves an interaction of the Θ and α subunits while in species that also have a polC homolog proofreading involves only the α subunit. This difference is probably sufficient to result in the general difference in mutation pattern but it is also the case that the Firmicutes show a stronger tendency to code genes on the leading strand than do other lineages [1]. This could conceivably contribute to an R-dependent mutation effect if there is a transcription-coupled repair system or if the lagging strand is more frequently in the single stranded state as a result, given the much greater tendency of ssDNA than dsDNA to undergo cytosine deamination [1].

If we consider overall skew (Table 6) there are some general points that are apparent, although these might partly reflect the non-random sample of genomes that have been sequenced. The high degree of overall skew in Borrelia noted by Loby and Sueoka [7] is apparent in our data, and the other taxa they noted as having strong skew (Treponema pallidum, Chlamydia muridarum and C. trachomatis) rank in the 100 chromosomes with strongest skew (see Additional file 1). The Alphaproteobacteria (particularly the genera Ehrlichia, Gluconobacter, Bartonella and Brucella) and the Firmicutes are disproportionately represented in the high skew genomes (both make up 35% of the 20 chromosomes with highest skew while Alphaproteobacteria and Firmicutes are 15.6% and 22.4% of the sequenced chromosomes respectively) while the Cyanobacteria tend to have low skew (7 of the 17 sequenced Cyanobacteria are in the 20 chromosomes with lowest skew). There are exceptions in each case (e.g. the endosymbiont Alphaproteobacterium Wolbachia and the Mycoplasma species) suggesting that mutational biases can vary dramatically and, perhaps, rapidly across lineages. In a previous survey, Rocha [1] also reported a low general skew in Cyanobacteria which is consistent with our finding, but also reported a low skew in the Alphaproteobacteria. The difference in the later case could be due to our use of an estimate from putatively neutral sites instead of a general composition bias.

**Conclusion**

The bipartition model allows us to quantify the contribution of the mutational difference between leading and lagging strands to nucleotide skew and also allows us to estimate the locations of the origin and terminus in chromosomes with bi-directional replication when the accumulation of skew is sufficiently strong. The model has

---

**Figure 3**

**Strand asymmetry in Firmicutes.** The same plot as in Figure 2 with the Firmicute chromosomes indicated as open points. Open squares indicate those of the class Mollicutes (which includes the Mycoplasma genus) while the open circles indicate all other Firmicutes.
several advantages over other methods of analyzing genome skew. First, using the model we can quantify the role of mutation in generating skew so that the effect on composition, for example codon bias, can be assessed. Second, it provides an objective method for locating origin and terminus that exploits composition bias. Finally, the model has the potential to be utilized in a maximum likelihood framework in order to analyze various aspects of genome structure, such as the effect of chromosome rearrangements on nucleotide composition.

Methods

Chromosomes analyzed
Primary chromosomes from 350 bacterial strains at the NCBI Genome Project website [38] as of June 20 2006 were downloaded as were 29 secondary chromosomes chosen from these species on the basis of annotation. Given the evidence for multiple origins of replications in Archaeabacteria chromosomes [39-41] the 27 chromosomes, which include one secondary chromosome, annotated as Archaea were removed to give a total of 352 chromosomes studied. Of this total, 5 chromosomes are linear and all others are circular. (Two chromosomes were incorrectly annotated by the NCBI file as linear; Baumannia cicadellinicola str. Hc NC_007984 [42], and Staphylococcus aureus susp. aureus NCTC8325 NC_007795 [43]. Linear chromosomes were analyzed in the same manner since they can also be divided into two replication arms. The taxonomic distribution of these 352 chromosomes is given in Table 1 and summary data are provided (see Additional file 1).

The bipartition model
We first develop a bipartition model that is the basis for ML methods. After deriving the core equations of the model we will discuss two specific applications of the model; an inference of the replication origin and terminus and an analysis of mutational contribution to skew. The approach of [7] to estimate the contributions of selection and mutation to skew is a special case of our model (see Equation 1 where $\pi_i$ represents the contribution of the R-independent and $\sigma_i$ the R-dependent factors respectively.

$$f_i = \pi_i + \sigma_i$$

Although it is not possible to calculate $\pi$ and $\sigma$ separately for each nucleotide, this division does allow us to estimate R-dependent and R-independent contributions to overall compositional skew (A-T and G-C differences). A circular chromosome that is replicated bi-directionally from a single origin is divided into two replication arms, each with its leading and lagging strands. We designate the annotated strand as + and define P as the arm in which this strand is the leading strand and $\rho$ as the arm within which it is the lagging strand. The R-dependent factors within P affecting the + strand are given by $\sigma^L$, while within $\rho$ they are given by $\sigma^L$ where L and $\rho$ refer to the leading and lagging strand respectively. Since $\sigma^L$ and $\sigma^R$ are complementary (for example, $\sigma^L = \sigma^R$), R-dependent components can be represented throughout the genome using only the $\sigma^L$ parameters, which from hereon will be written without a superscript designation. We can then write the frequencies of nucleotide $i$ on the + strand within a replication arm (i.e. $P_i$ or $\rho_i$) as in Equations 2a and 2b where $i$ refers to any of A, C, G or T and $j$ to the complementary nucleotide. Note that the R-dependent parameter ($\sigma^L$) changes to the complementary nucleotide ($\sigma^R$) between leading and lagging strands.

$$P_i = \pi_i + \sigma_i$$

$$\rho_i = \pi_i + \sigma_j$$

Given our partition of substitution dynamics, the A+T composition ($C_{A+T}$) of the + strand within either replication arm can be written as Equation 3, with a similar equation for $C_{G+C}$ ($C_{G+C} = 1 - C_{A+T}$).
The skew parameters are given in Equations 4a and 4b where \( \pi_{ij} \) represents the R-independent and \( \sigma_{ij} \) the R-dependent components of compositional skew.

\[
\pi_{AT} = \pi_A - \pi_T \quad \pi_{GC} = \pi_G - \pi_C \\
\sigma_{AT} = \sigma_A - \sigma_T \quad \sigma_{GC} = \sigma_G - \sigma_C
\]

(4a) \hspace{1cm} (4b)

In the absence of selection and transcription-coupled effects, \( \pi_{ij} = 0 \) and \( \sigma_{ij} \) represents the skew generated by mutational differences between the leading and lagging strands.

We can now express the nucleotide frequencies for the + strand from Equation 1 in terms of \( C_{A+T} \) and the skew generated by mutational differences between the leading and lagging strands.

The skew parameters are given in Equations 5a through 5d and similar equations hold for \( \rho \).

\[
P_A = \frac{1}{2}(C_{A+T} + \pi_{AT} + \sigma_{AT}) \quad (5a) \\
P_T = \frac{1}{2}(C_{A+T} - \pi_{AT} - \sigma_{AT}) \quad (5b) \\
P_C = \frac{1}{2}((C_{G+C} - \pi_{GC} - \sigma_{GC}) \quad (5c) \\
P_G = \frac{1}{2}((C_{G+C} + \pi_{GC} + \sigma_{GC}) \quad (5d)
\]

The values \( C_{A+T}, \pi_{AT} \) and \( \sigma_{AT} \) can now be rewritten as in Equations 6a through 6c.

\[
C_{A+T} = \frac{1}{2} \left[ (P_A + P_T) + (\rho_A + \rho_T) \right] \quad (6a) \\
\pi_{AT} = \frac{1}{2} \left[ (P_A - P_T) + (\rho_A - \rho_T) \right] \quad (6b) \\
\sigma_{AT} = \frac{1}{2} \left[ (P_A - P_T) - (\rho_A - \rho_T) \right] \quad (6c)
\]

The parameters for GC content and skew \( C_{G+C}, \pi_{GC} \) and \( \sigma_{GC} \) can be calculated in a similar manner. These equations (6a, 6b and 6c) represent the core of our model for a chromosome with two replication arms since they allow us to estimate the R-independent (\( \pi \)) and R-dependent (\( \sigma \)) components of skew from nucleotide composition within each replication arm. Note that although we have been discussing circular chromosomes, the model also applies to linear chromosomes with bi-directional replication from a single origin. This is the general model that is implemented below in ML applications.

**Classifying sites to minimize effects of selection**

As already noted, selective pressures for amino acid composition can contribute to skew, particularly if there is a coding strand bias in the genome [8]. Since these pressures are part of the R-independent component, if we want to measure the R-dependent component of skew then we will need to apply our model to neutral sites. This is possible since Equations 5 & 6 can be applied to any subset of sites within a chromosome. Therefore, if we use only the composition of relatively neutral sites, such as fourfold degenerate sites or intergenic sites, we can use the parameter \( \sigma \) determined from them as an estimate of the R-dependent component of skew. Nucleotide sites in each genome were classified based on the NCBI annotation according to codon position \( (C_1, C_2, C_3 \) for CDS-coding genes) on both + and - strands, as RNA-coding or as intergenic (IG). Third codon position sites were further subclassified as \( D_j \) if four-fold degenerate (this did not include those that are from a codon coding a six-fold degenerate amino acid).

**Maximum likelihood Implementation**

Equations 6a-c (and similar ones for GC) allow an estimation of model parameters directly from nucleotide composition parameters within replication arms. The model can also be implemented in a ML framework using a variety of specific methods, discussed separately below, that differ in the constraints they introduce. A comparison of different methods will allow us to assess the explanatory power of the constraints. To implement these ML methods, the genome is divided into two replication arms. (This can be achieved either by using an annotated [Ori, Ter] or the putative locations determined using the method we will describe below.) This genome division defines the number of sites of each type within each replication arm such that there are \( M_k \) sites of type \( k \) that are nucleotide \( i \) in region \( P \).

Expected nucleotide frequencies are calculated from the model parameters according to equations 5a-d (with corresponding equations for the \( \rho \) replication arm). We can then calculate the likelihood of the model correctly predicting the observed nucleotide composition of both replication arms. For site type \( k \) within the \( P \) replication arm, this likelihood is given by Equation 7a.

\[
L^k_P = \prod \left( p^k_i \right)^{M^k_i} \quad (7a)
\]

The likelihood for the replication arm \( P \) is then the product of the likelihood of each site type (Equation 7b). \( L_P \) can be calculated in the same manner and the overall likelihood is the product of \( L_P \) and \( L_P \).

\[
L_P = \prod L^k_P \quad (7b)
\]

**Method \( M_{obs} \)**

Model parameters for each site type are calculated according to Equations 6a-c (and corresponding equations for
GC) using the observed nucleotide frequencies within each replication arm. The likelihood is not maximized. No constraints are introduced but sites are classified into 7 types: intergenic (IG), first codon position (C1), second codon position (C2) and third codon position (C3). The latter three are further divided into + and - strand (+ strand being defined by the NCBI file) to yield C1+, C1-, C2+, C2-, C3+ and C3-. RNA coding sites are ignored, as are sites that are ambiguous in the NCBI annotation.

**Method M0**

This method uses the same approach and site classification as Mobs. An initial guess for site parameters was obtained from Mobs and maximum likelihood parameters were obtained using a simplex algorithm [44]. Each site type is optimized independently. There are 5 DF for each site type giving a total of 35 DF.

**Method M1**

This method introduces the constraint that the π parameters calculated for coding sites on the two chromosome strands, such as for example C1+ and C1-, be complementary (\(\pi_{ij} = -\pi_{ji}\)). This constraint allows us to assess whether or not there is a significant difference between coding sites on the two strands using a likelihood ratio test. Thus there is only a single set of 5 parameters for each of the 3 CDS site types and 5 for IG sites, yielding 20 DF. An initial guess for these 20 parameters was obtained from Mobs and then the values that maximized the total likelihood were obtained using the simplex algorithm.

**Method M2**

This method has the same constraints as the M1 method with the additional constraint that the σAT and σGC parameters are equal across all site types. Therefore, if mutational biases are consistent across sites, then comparing this method to an unconstrained method allows us to assess whether or not selection significantly affects our measurement of the R-dependent component of skew. There are 14 DF in this method (3 for IG and each of the 3 CDS types, plus a single \(\sigma_{AT}\) and \(\sigma_{GC}\) for all site types). Parameters were calculated as described for M1.

**Method M3**

This method removes the site classification such that all nucleotide sites in the chromosome are assumed to be equivalent. It retains the constraint that parameters for the two strands be complementary. This allows us to assess the value of site classification. This method has 5 DF and values were obtained as described for M1.

**Application to the E. coli K12 chromosome**

The effects of constraints using the ML methods will be illustrated by application to the E. coli K12 chromosome. We divided this chromosome into two replication arms based on the annotated origin/terminus sites at approximately 3,924,000 and 1,589,000 respectively in the NCBI file and used Equations 5–7 for C1+, C1-, C2+, C2-, C3+ and intergenic (IG) sites to determine the likelihood of method Mobs. The likelihood obtained from mean nucleotide composition within each replication arm (Mobs) does not differ appreciably from that made according to ML (Table 7). Methods M1 and M2 produce significant decreases in likelihood, indicating that selection significantly affects our measure of the R-dependent component of skew.

The R-independent parameters (\(\pi\)) for coding sites show the expected complementarity between the + and - strands (Table 8), meaning that the \(\pi\) value calculated for the sites coded on one strand is the negative value of the \(\pi\) value for the sites coded on the other strand (since both parameters are calculated from the composition of the strand given in the NCBI file). The exception is the C2 site class for which the two strands show significantly different estimates of \(\pi_{AT}\), even accounting for the complementarity, suggesting a different average composition of proteins coded on the two strands. For IG sites, the \(\pi_{AT}\) and \(\pi_{GC}\) parameters are not significantly different from zero indicating that there is no net R-independent skew across the genome. This lack of an R-independent skew effect on IG sites, however, does not necessarily indicate a lack of selection; since we cannot assign a strand to intergenic sites, any selective effect could be equally distributed across the two strands with the result that we observe no net skew.

There is a significant R-dependent effect (\(\sigma\)) for IG sites. The \(\sigma\) values are very similar for + vs - strands within the C1 and C3 site types, but second position codon sites (type C2) have significantly different estimates of \(\sigma_{AT}\). This difference indicates that selective constraints on the sites of the C2 class are not distributed equitably across the two leading strands of the genome. As with the different \(\pi\) values at C2 sites this suggests that there is a difference in protein composition between + and - strands, but it also suggests that there is a preference for coding certain types

**Table 7: Likelihood comparisons for the different methods when applied to the E. coli K12 chromosome CDS and intergenic sites**

| Method | Log (L)   | -2 × Diff1 | Probability (DF1, DF2)2  |
|--------|-----------|------------|--------------------------|
| Mobs   | -2696055.16 | NA         | NA                       |
| M0     | -2696052.54 | reference  | NA                       |
| M1     | -2696075.49 | 45.9       | 5.5 × 10⁻⁵ (35, 20)      |
| M2     | -2696239.16 | 373.2      | < 10⁻⁶ (35, 14)          |
| M3     | -2740719.34 | 89,334     | < 10⁻⁶ (35, 5)           |

1Relative to M0.

2Probability of the chi-square test with (DF1 - DF2) degrees of freedom.
of genes (that differ in average composition) on the leading strand.

Overall, the results from this analysis of *E. coli* provide two specific points that are relevant to the analysis of other chromosomes. First, the method with the fewest constraints (M0) provides the best fit so we will use this method in our large-scale analyses. Second, given the influence of selection on estimates of \( \sigma_{AT} \) and \( \sigma_{GC} \), we must limit the analysis to relatively neutral sites in order to separate the mutational contribution from selection. Two possible choices are intergenic (IG) sites and four-fold degenerate (D4) CDS sites and both of these will be utilized and compared in our applications of the bipartition model.

### Estimating the R-dependent component of composition skew

Lobry and Sueoka [7] introduced a graphical method for estimating the contributions of selection and mutation to composition skew. This involves plotting the T/A skew (i.e., \([-f_{T} - f_{A}]/[f_{T} + f_{A}]\)) against the G/C skew for third codon (i.e., putatively neutral) sites of each gene in a genome and calculating mid-points for leading and lagging strand genes. The length of the line connecting these two points (their \( B_{II} \) parameter) is an estimate of the role of mutational bias to skew while the distance from the (0, 0) point to the midpoint of this connecting line (their \( B_{I} \) parameter) is an estimate of the contribution of selection [7].

The parameters derived from our model can be used to estimate these two parameters as summarized in Equations 8a and 8b. They only require the determination of the unordered pair \([S_{1}, S_{2}]_{\text{ML}}\) since \( B_{I} \) and \( B_{II} \) are invariant if the origin and terminus are interchanged.

\[
B_{I} = \text{SQRT} \left( \frac{\sigma_{AT}}{C_{AT}} \right)^{2} + \left( \frac{\sigma_{GC}}{C_{GC}} \right)^{2} \quad (8a)
\]

\[
B_{II} = \text{SQRT} \left( \frac{\sigma_{AT}}{2C_{AT}} \right)^{2} + \left( \frac{\sigma_{GC}}{2C_{GC}} \right)^{2} \quad (8b)
\]

However, there are some disadvantages to this graphical approach that our model can improve on. One is that it weights genes equally regardless of length, another is that the statistical confidence of the parameters is unclear as is the biological meaning of the values obtained. Most importantly, it does not allow us to estimate the contribution of mutation to G/C and T/A skew separately. Using the bipartition model we can make a direct estimate of the relative contributions of R-dependent (\( \sigma \)) effects to the composition of G and T (since a keto skew is most commonly observed) for each type of site that can be classified. The measures we propose are \( R_{C} \) and \( R_{T} \) as given in Equations 9a and 9b.

\[
R_{C} = \frac{\sigma_{GC}}{C_{GC} - \pi_{GC}} \quad (9a)
\]

\[
R_{T} = -\frac{\sigma_{AT}}{C_{AT} - \pi_{AT}} \quad (9b)
\]

At selectively neutral sites, and assuming that there are no transcription-coupled effects, these two values represent the fractional increase (or decrease) in the G and T compositions on the leading strands as a result of a difference between leading and lagging strand mutation bias. Thus, the model provides a statistical framework to assess significance and to easily compare mutational effects across genomes. We calculated \( R_{C} \) and \( R_{T} \) for each bacterial chromosome from Equations 9a and 9b (on the leading strand) based on the assignation of \([P_{D}, P_{J}]_{\text{ML}}\) as described. To use sites that we thought would be relatively neutral we used model M0 parameters for intergenic (IG) and for C3 sites that coded for a four-fold degenerate amino acid (D4). Since parameters for D4 sites were determined separately for both plus and minus strands the values from the two strands were averaged for an overall estimate of the R-dependent effect.

---

**Table 8: Model parameters for the *E. coli* K12 chromosome when the M0 method is implemented**

| Site Class | \( C_{AT} \) | \( \pi_{AT} \) | \( \sigma_{AT} \) | \( \sigma_{GC} \) | \( \pi_{GC} \) | \( \sigma_{GC} \) |
|-----------|-------------|-------------|-------------|-------------|-------------|-------------|
| IG        | 0.578       | 0.0012      | -0.0053     | 0.0015      | -0.0004     | 0.0180      |
| C1*       | 0.411       | 0.0944      | -0.0023     | 0.104       | 0.0102      | 0.0072      |
| C2*       | 0.409       | -0.0954     | -0.0010     | 0.107       | -0.1088     | 0.00109     |
| C3*       | 0.593       | -0.0168     | 0.0079      | 0.0461      | 0.0048      | 0.0045      |
| C4*       | 0.594       | 0.0133      | -0.0019     | 0.0449      | 0.0043      | 0.0038      |
| D4*       | 0.442       | -0.0782     | -0.0056     | 0.0219      | 0.0203      | 0.0249      |
| D4*       | 0.439       | 0.0013      | -0.0064     | 0.0196      | 0.0021      | 0.0241      |

1 Sites are classified as intergenic (IG) or coding (C) with the latter further classified by codon position, as indicated by the subscript, and chromosome strand, as indicated by the superscript. D4 are four fold degenerate C3 sites.

2 The ML value is given with the 95% interval given in brackets.
Estimation of parameter uncertainty
The 95% uncertainty range in model parameters and in R_p and R_c was estimated using the Metropolis-Hastings rejection sampling Monte Carlo algorithm [45,46] as follows. First, the ML vector of parameters, with Likelihood L_0, was determined by the simplex algorithm. Next, one parameter was altered using a Gaussian random number with mean 0 and standard deviation σ which is input arbitrarily to start. The Likelihood for the new vector, L_1, was calculated and the new vector “accepted” if L_1 > L_0 or with probability L_1/L_0 if L_1 < L_0. A burn-in phase involved repeating the acceptance process in sets of 100 with successively decreasing σ values until the average acceptance rate of new vectors is 50% over the 100 trials. After the burn-in was complete, the final burn-in vector of parameters and σ value were used to generate 10,000 parameter vectors by altering parameter values as above. The 10,000 were sorted for each parameter and the 95% range determined from the sorted set.

Application to the E. coli K12 chromosome
Estimates of R-dependent effects, R_c and R_p were calculated for D_4 and IG sites from the data in Table 8. For D_4 sites in E. coli K12 the M_0 model gives R_c = 0.0716 (+ strand = 0.0778; - strand = 0.0653) and R_p = 0.0215 (+ strand = 0.0181; - strand = 0.0248), while for IG sites the values are R_c = 0.0428 and R_p = 0.009. These results indicate that R-dependent effects increase the content of G at D_4 sites on the leading strand by about 7% relative to the composition in the absence of R-dependent effects (decreasing C by the same amount) and the content of T at these sites by about 2%.

ML determination of replication arms based on composition skew
The bipartition model can be used to infer the replication arms in genomes with a replicochrome structure by maximum likelihood since it provides a probability for each of the four nucleotides. This will be important for applications of our model to genomes for which an origin and terminus have not been annotated but it also introduces a more formal approach than skew plots to estimating these two loci.

Two chromosomal locations, S_1 and S_2 (which need not divide the chromosome into equal halves), define two replication arms, P and p, and thus determine the number and kind of sites within each. The likelihood of such a division is calculated according to model M_obs for C_1 [+/], C_2 [+/], C_3 [+/], and IG sites. The likelihood for all possible unordered [S_1, S_2] pairs determines the global ML (designated here as [S_1, S_2]ML). Ideally we would calculate the likelihood for every possible [S_1, S_2] division in order to determine [S_1, S_2]ML but given computational constraints we utilized the following strategy. A rough, initial likelihood map is made by computing log likelihood values over a 50 × 50 grid on the [S_1, S_2] plane. The maximum of this rough map was then used as the initial point to maximize the likelihood using the same simplex algorithm used for parameter optimization [44]. From [S_1, S_2]ML we calculate the chromosome division statistic C_d by Equation 10, in which min(f_1, f_2) is the minimum of the two chromosome fractions generated by S_1ML and S_2ML. The C_d statistic can range from 0, in cases where there is equitable division of the genome, to 1, in cases where the origin and terminus are at the same chromosome location yielding essentially a single replication arm.

A visualization of the ML surface and estimates of uncertainty in C_d, S_1 and S_2 (as well as model parameters, see below) were obtained by the Metropolis-Hastings rejection sampling Monte Carlo algorithm, using a random walk from [S_1, S_2]ML. Proposal parameter (S_1 and S_2) values were obtained by a normally distributed random step of zero mean and a standard deviation that was chosen after a burn-in series to give an approximately 50% acceptance rate. If the proposed parameters improved the likelihood they were accepted. Otherwise, they were accepted with a randomly generated probability equal to the likelihood ratio. Ten thousand accepted parameter sets were sorted to give a 5-to-95 percentile range.

Application to the E. coli K12 chromosome
The replication origin for E. coli K12 has been located empirically and provides a good demonstration and test of our ML implementation. The log likelihood surface (see Figure 4) for the E. coli K12 chromosome (NC_000913) shows the two symmetrical peaks that differ in assignment of leading vs lagging strands as expected for the bipartition model. S_1ML was identified (correctly) as the putative origin by the distribution of rRNA sites. Thus, |P_O|, P_p^ML = [0.842, 0.342], where the locations are given as a fraction of the chromosome length (4,639,675) starting at the NCBI site 1. These two fractions represent the approximate nucleotide positions 3,908,000 and 1,589,000. The 95% intervals are [0.828 – 0.871, 0.334 – 0.352] for PO and P_p^ML. The oriC and terminus (dif site) locations in E. coli are annotated at roughly positions 0.846 (3,924,000) and 0.342 (1,589,000) respectively in the NCBI file. The annotated origin and terminus both lie within the confidence interval and are very close to the ML estimates. Therefore, the bipartition model provides an accurate estimation of the origin of replication in this chromosome. Four additional examples are provided as supplemental material (see Additional file 2).
Identification of the putative origin and terminus of replication

The likelihood surface has two peaks because exchanging $S_1^{ML}$ and $S_2^{ML}$ yields complementary $\sigma$ parameters but the same likelihood. Thus, independent information must be used to determine a putative origin and terminus pair. We assigned the origin in one of several ways. For the 5 linear chromosomes we made use of the annotated origins in the NCBI files, which are all located near the center of the chromosome, by assigning whichever of $S_1^{ML}$ or $S_2^{ML}$ was closest to the annotated origin. For circular chromosomes we used annotated ribosomal RNA (rRNA) genes, where available, which are organized such that they are predominantly on the leading strand of replication across microbial genomes [22]. A third method for assigning the origin was utilized in the 14 chromosomes for which the rRNA annotation was not available and the 8 chromosomes for which there was an equal division of rRNA genes between the leading and lagging strands. In these cases we assigned the leading strand in each replication arm as the strand for which $\sigma_{GC} > 0$ at fourfold degenerate sites ($\sigma$ is defined for the leading strand). This use of $\sigma_{GC} > 0$ as the criterion is based on the identification by Lobry and Sueoka [7] that, in the vast majority of bacterial chromosomes they studied, the leading strand has a skew of G > C. This last approach means that we are limited to knowing whether or not we find a G&T or G&A bias on one strand (and thus C&A or C&T on the other) in these particular chromosomes without definitively assigning the skew to leading or lagging strand. Regardless of the method used, the result is that we define either $S_1^{ML}$ or $S_2^{ML}$ as the putative origin, now designated $P_O^{ML}$, and the other site as the putative terminus, now designated $P_T^{ML}$, to yield the ordered pair of sites $[P_O, P_T]^{ML}$. This provides a formal approach that is preferable to estimates based on graph-based approaches [16-19] particularly in cases where there is a small amount of skew that can be difficult to study by eye. Additionally, the model allows us to generate confidence intervals by a Monte Carlo method as outlined above for the ML surface.

Software

All analyses were performed using C source code programs written and compiled for Mac OSX by RAM. A description of the three core programs used to determine model M0 parameters, the ML [Ori, Ter] location and plot a visualization of the ML surface is provided as additional material (see Additional file 3). Code is available upon request from morton@mcmaster.ca.

List of abbreviations

ML – maximum likelihood, IG – intergenic sites, CDS – nucleotide sites in a protein-coding DNA sequence, divided into $C_i$ for the $i^{th}$ codon position and further into $D_4$ for four-fold degenerate $C_3$ sites, R-(in)dependent – Replication arm-, DF – degrees of freedom.

Competing interests

The author(s) declares that there are no competing interests.

Authors’ contributions

RAM wrote the computer programs to implement the model. All other contributions were equal. All authors read and approved the final manuscript.
Additional material

Additional file 1
Summary information and data for 352 bacterial chromosomes. Data, including the NCBI accession numbers, for each of the bacterial chromosomes used in this study. Click here for file [http://www.biomedcentral.com/content-supplementary/1471-2164-8-369-S1.txt]

Additional file 2
Additional examples of likelihood surfaces together with [G-C] and [A-T] skew walks. Figure 1. E. coli [NC_000913]. Figure 2. S. melliloti 1021 plasmid pSymA [NC_000307]. Figure 3. Nostoc sp. PCC 7120 [NC_0033272]. Figure 4. Prochlorococcus marinus str. MIT 9313 [NC_005071]. Click here for file [http://www.biomedcentral.com/content-supplementary/1471-2164-8-369-S2.pdf]

Additional file 3
Description of computer programs. A brief summary of the code for programs used to determine 1) model M0 parameters, 2) the ML [Ori, Ter] location and 3) plot a visualization of the ML surface. Click here for file [http://www.biomedcentral.com/content-supplementary/1471-2164-8-369-S3.doc]

Acknowledgements
We would like to thank three anonymous reviewers for helpful comments.

References
1. Rocha EP: The replication-related invasion of bacterial genomes. Microbiology 2004, 150:1609-1627.
2. Bird RE, Louarn J, Martuscelli J, Caro L: Origin and sequence of chromosome replication in Erichesia coli. J Mol Biol 1972, 70:549-566.
3. Marians KJ: Prokaryotic DNA replication. Annu Rev Biochem 1992, 61:673-715.
4. Song J, Ware A, Liu SL: Wavelet to predict bacterial ori and ter: a tendency towards a physical balance. BMC Genomics 2003, 4:17-30.
5. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y: The Complete Genome Sequence of Escherichia coli K-12. Science 1997, 277:1456-1462.
6. McLean MJ, Wolfe KH, Devine KM: Base composition skews, replication orientation and gene orientation in 12 prokaryote genomes. J Mol Evol 1998, 47:691-696.
7. Lobry JR, Louarn JM: Polariisation of prokaryotic chromosomes. Curr Opin Microbiol 2003, 6:101-108.
8. Karlin S: Bacterial DNA strand compositional asymmetry. Trends Microbiol 1999, 7:305-308.
9. Francino MP, Ochman H: Strand asymmetries in DNA evolution. Trends Genet 1997, 13:240-245.
10. McNerney JO: Replication and transcriptional selection on codon usage in Borrelia burgdorferi. Proc Natl Acad Sci USA 1998, 95:10698-10703.

Additional material
Page 16 of 17 (page number not for citation purposes)
Genome sequence of Halo bacterium species NRC-1. Proc Natl Acad Sci USA 2000, 97:12176-12181.

31. Giovannini SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbo L, Eads J, Richardson TH, Noordewier M, Rappe MS, Short JM, Carrington JC, Mathur EJ. Genome streamlining in a cosmopolitan oceanic bacterium. Science 2005, 309:1242-1245.

32. Guy L, Karamata D, Moreillon P, Rosen CAH. Genometrics as an essential tool for the assembly of whole genome sequences: the example of the chromosome of Bifidobacterium longum NCC2705. BMC Microbiology 2005, 5:60-69.

33. Zakrzewska-Czerwinska J, Schrempf H. Characterization of an autonomously replicating region from the Streptomyces lividans chromosome. J Bacteriol 1992, 174:2688-2693.

34. Goodner B, Hinkle G, Gattung S, Miller N, Blanchard M, Qurrollo B, Goldman BS, Cao Y, Askenazi M, Halling C, Mullin L, Houmiei K, Gordon J, Vaudin M, Iartchouk O, Epp A, Liu F, Wollam C, Allinger M, Doughtry D, Scott C, Lappas C, Markelz B, Flanagan C, Crowell C, Gurson J, Lomo C, Sear C, Scrub G, Cielo C, Slater S. Genome sequence of the plant pathogen and biotechnology agent Agrobacterium tumefaciens C58. Science 2001, 294:2323-2328.

35. Bignell C, Thomas CM. The bacterial ParA-ParB partitioning proteins. J Bacteriol 2001, 191:1-34.

36. Ogawara N, Yoshikawa H. Genes and their organization in the replication origin region of the bacterial chromosome. Mol Microbiol 1992, 6:629-634.

37. Andolfatto P. Adaptive evolution of non-coding DNA in Drosophila. Nature 2005, 437:1149-1152.

38. NCBI Genome Project [http://www.ncbi.nlm.nih.gov/genomes].

39. Lundgren M, Andersson A, Chen L, Nilsson P, Bernander R. Three replication origins in Sulfolobus species: Synchronous initiation of chromosome replication and asynchronous termination. Proc Natl Acad Sci USA 2004, 101:7046-7051.

40. Lundgren M, Bernander R. Archaeal cell cycle progress. Curr Opin Microbiol 2005, 8:662-668.

41. Kelman Z, White MF. Archaeal DNA replication and repair. Curr Opin Microbiol 2005, 8:669-676.

42. Wu D, Daugherty SC, Van Aken SE, Pai GH, Watkins KL, Khouri H, Tallon LJ, Zaborsky JM, Dunbar HE, Tran PL, Moran NA, Eisen JA. Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. Proc Biol Sci 2006, 273:1079-1092.

43. Pattee PA. Chromosomal map location of the alpha-hemolysin structural gene in Staphylococcus aureus NCTC 8325. Infect Immun 1986, 54:593-596.

44. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes in C. 2nd edition. Cambridge University Press; 1992:41.

45. Hastings WK. Monte Carlo sampling methods using Markov Chains and their applications. Biometrika 1970, 57:97-109.