בנחתה של תופעת שמידות גני הבקריה במינו

שונם של בקריה

היבור והווש חקל מועדירשת קבלת ההנהרת
"מוסמך אוניברסיטה" - M.Sc. באוניברסיטה תל אביב
ביח"ש בלטנשיק למדעי המחשב

עד ד"ר
אורן איש-עופי

עבודת המחקר בוצעה בהדריכה של פרופסור איתן רופין

תשרי תשע"ד
A cross-species study of the evolutionary conservation of essential genes in bacteria

This thesis is submitted in partial fulfillment of the requirements towards the M.Sc. degree

Tel-Aviv University
The Blavatnik School of Computer Science

by

Oren Ish-Am

The research work in this thesis has been carried out under the supervision of Prof. Eytan Ruppin

September, 2013
Acknowledgments

First, I would like to express my sincere gratitude to my supervisor Professor Eytan Ruppin, for setting high standards and making me adhere to them in his charming and yet uncompromising manner; for supporting my work and unorthodox ideas, always embracing creativity and finding pith and merit in every change of fortune.

I would like to thank my lab members: Keren Yizhak, Mathew Oberhardt, Yoav Teboulle, Livnat Jerby, Yedael Waldman, Allon Wagner and Raphi Zarecki for their willingness to share both time and knowledge and for their invaluable advice.

The research described in this thesis was funded by a generous support from the Edmond J. Safra Bioinformatics Program at Tel Aviv University.

Finally, I would like to thank my dear wife Adi, whose belief in education and love for academia, prompted me into my master's degree and who supported me through the uncertainties of research and the long toils of study.
תקציר

אחת המנחות החשובות במיקרוביולוגיה היאشنאיות של הגנים המבクラスים פונקציונליות זרים.

לאחר שנות בקבי איטי, מונחים שlessnessות פחותות בגנים הכרחיים, ולא ניתןлежותן גורמות למ(shift, עמעון, שיפורים)

האגרונומיים, הניבים של הגנים פונקציונליים מתוכחת בברר ומתחם של מחקרי רֶב. עד ימינו,

העדות לברק של שני מונחים בקבי איטי גנים מפגינים אינד משמעויות.chod המכסולים

הحذر של שגוי זה הוא יותר見る לתו גנים הכרחיים אלא. רווח מחקרים הקהולים בכמה

על גנים הכרחיים שנועדו ביוסי, תובנה על ממשקן, שיקולים יוני לתמי תיוד ההבדלים

של הבקריה מטול בכפה. קיינים מחקרים של הגנים הכרחיים עד מחזור ת שעל יוני

מטבוליזם, מהווה השפעת במדועה ברקודה שבחזה אם ניצים את המחקר הראשים.

השעון בשתיו הרקמה הפיתולית, המסורה מצא הם לריק קר dob המכסולים.

אנו מציינים את מחקרים חדשנ—I. מחקרים של המחקר המבשלות – ECVA, האומפר של מחקרו את הגנים של שומן ב-

הגרו מטול ברק. לא

генומיה על perí לכל מץ אשק.

אנו מודחים על שלושה מציאת מחויר: 1) בקריה מטול לישומות של הגנים הכרחיים לועמון relevance

הגרו מטול ב-2) הגנו דו תחילת образ, את הגנים מפרשות על פי פסקטייה רב. 3) לא

מציאת פונטייה בולוונ (כוללים בתקי, סיבוב נודל, פילוגונטיקה) המפריד בין בקריה לעניין

שומן שומן, המ מחויר OWNER בשום שלמה פיתוחה ומיון.
1 Abstract

One of the basic postulates of molecular evolution is that functionally important genes should evolve at a slower pace than genes of lesser significance. Essential genes, whose knockout leads to a lethal phenotype, are considered of high functional importance and have been thoroughly investigated in many studies. Yet, whether they are truly more conserved than nonessential genes has been the topic of much debate, fuelled by a host of contradictory findings.

A key difficulty in rigorously addressing this question has been that the answer depends on determining which genes are essential and which are not. Previous studies have mostly looked at essential genes experimentally determined on synthetic lab media, whose relation to the growth conditions of bacteria in their natural habitats is highly questionable. As the identity of essential genes is strongly media-dependent (especially, but not exclusively metabolic genes), this presents a major shortcoming. Here we conduct the first large-scale systematic study addressing this issue, utilizing genome-scale metabolic modeling and encompassing a wide range of bacterial species. We present a novel computational approach, *Evolutionary Conservation Variation Analysis* (ECVA), which enables us to examine the range of essential vs. nonessential gene conservation *across all possible media* and classify bacterial species into five evolutionary conservation categories. Our key findings are threefold: (1) bacterial species show a tendency for differential conservation of their essential genes, however (2) this tendency is highly variable and media-dependent. (3) We fail to find significant differences between central biological traits (such as growth rate, lifestyle and phylogeny) between bacteria belonging to the five different evolutionary conservation categories, leaving the source of this variance an open, interesting biological question.
Abstract .......................................................................................................................... 5

Introduction .................................................................................................................... 8
  2.1 Gene Essentiality and Evolutionary Conservation ..................................................... 8
  2.2 Genome-Scale Metabolic Modeling (GSMM) ........................................................... 11
  2.3 Metabolic Models Used in Study ............................................................................ 17
  2.4 Evolutionary Conservation of Essential Genes in Bacteria .................................... 17

Methods .......................................................................................................................... 19
  3.1 Biological Attributes ............................................................................................... 19
  3.2 dN/dS Estimates ........................................................................................................ 20
  3.3 Metabolic Model Processing .................................................................................... 20
  3.4 Calculating Essential Genes .................................................................................... 25
  3.5 AE, PE, APE and NE genes ..................................................................................... 25
  3.6 Definition of KOR and non-KOR Split ..................................................................... 27
  3.7 Correcting for Multiple Hypotheses ......................................................................... 30

Results ............................................................................................................................. 32
  4.1 Overview .................................................................................................................. 32
  4.2 ECVA and ECOEDS ............................................................................................... 33
  4.3 Evolutionary Conservation Variation Analysis (ECVA) ............................................ 36
  4.4 Are Bacterial Essential Genes more Evolutionary Conserved? ............................... 43
  4.5 KOR splits and Biological Attributes .................................................................... 44
  4.6 KOR Split and Phylogeny ...................................................................................... 48
  4.7 KOR Split and Habitat ............................................................................................ 49

Discussion ......................................................................................................................... 50

Supplementary Information ............................................................................................. 54
  6.1 Gene Essentiality Fast Calculations ....................................................................... 54
  6.2 APE Completeness .................................................................................................. 56
  6.3 ECOEDS SA and Purification Details .................................................................... 56
  6.4 EDC Media vs. Synthetic Media KOR Scores ......................................................... 58
  6.5 EDC Media vs. Random Media KOR Scores .......................................................... 60
  6.6 KOR classification of S. cerevisiae ......................................................................... 61
  6.7 Models database and scores ................................................................................... 61

Bibliography ..................................................................................................................... 63
2 Introduction

2.1 Gene Essentiality and Evolutionary Conservation

Individual genes within a given species genome contribute differently to the survival and propagation of the organism. According to their known functional profiles, genes can be divided into two categories: essential and nonessential genes. These are classic molecular genetic designations that relate to the functional significance of a gene with respect to its effect on organismic fitness (Jordan, Rogozin, Wolf, & Koonin, 2002). A gene is considered to be essential if its knockout results in (conditional) lethality or infertility. On the other hand, nonessential genes are those for which knockouts yield viable and fertile individuals. In light of their indispensable nature, the study of essential genes not only has great theoretical relevance in exploring the genesis of life but also plays an important practical role in modeling the design of antibacterial drugs (Gong et al., 2008).

Almost four decades ago, Alan Wilson and colleagues (Wilson, Carlson, & White, 1977) proposed that the genetic rate of evolution should be dependent on gene importance, according to which one would expect that essential genes should evolve more slowly than nonessential genes. This has been termed the knockout-rate (KOR) hypothesis (Hurst & Smith, 1999), linking between gene functional indispensability and rate of evolution.

Since the publication of the KOR hypothesis, the question of whether essential genes are in fact more evolutionary conserved than nonessential genes has been widely debated and results have been equivocal thus far. In eukaryotes, most studies refute the KOR hypothesis: Hurst and Smith (Hurst & Smith, 1999) measured the substitution rates between mouse and rat orthologous genes. They also tested for
protein importance using the knockout phenotypes of 175 mouse genes, and concluded that there was no difference in the evolutionary rate between essential and nonessential genes. Hirsh and Fraser (A E Hirsh & Fraser, 2001) observed that essential genes do not evolve slower than nonessential genes in *S. cerevisiae*. They did, however, find a weak negative correlation between gene importance and evolutionary rate when analyzing the evolutionary distances between the genes of yeast and the nematode *C. elegans*. Pál and Hurst (Pál, Papp, & Hurst, 2003) analyzed protein substitution rates using three close relatives of *S. cerevisiae* and concluded that it was gene expression levels and *not* essentiality that were responsible for the small effect of dispensability on protein substitution rates. In contrast, Hirsh and Fraser (Aaron E. Hirsh & Fraser, 2003) using a different methodology to identify orthologs and a different set of transcriptome data, confirmed their previous conclusions, showing a significant correlation between dispensability and evolutionary rate, even when controlled for expression levels. Zhang and He (J. Zhang & He, 2005) working with yeast, showed that protein evolutionary rate was significantly correlated to dispensability, even when gene expression levels were controlled for and duplicate genes excluded. Wang and Zhang (Wang & Zhang, 2009) studying *S. cerevisiae*, claim that the weakness of the correlation between gene importance and evolutionary rate is factual and does not result from the disparate data sources and analysis methods used by the previous studies. They also show that the principle of slower evolution of more important genes *does* have some predictive power when genes with vastly different evolutionary rates are compared.

In the bacterial domain, again, opinions differ. Jordan et al. (Jordan et al., 2002) found that, based on essentiality information from an experimental (null mutation)
database, the KOR hypothesis holds for *E. coli* and also to a lesser degree for *Helicobacter pylori* and *Neisseria meningitidis* (inferring essentiality by homology to *E. coli*). Rocha and Danchin (Rocha & Danchin, 2004) performed regression analysis checking the relative importance of protein expression levels, functional category, essentiality and metabolic costs of amino acid residues on a protein's evolutionary conservation in *E. coli* and *Bacillus subtilis*. They claim that in both bacteria, gene expression level was by far the most important driving force constraining amino acid substitution rate. Gong et al. (Gong et al., 2008) explored *E. coli* using the ERI measure to estimate the conservation of a protein (ERI is a value that depicts the propensity of a protein to have orthologs in other fully sequenced genomes, and reflects how persistent a gene is (Fang, Rocha, & Danchin, 2005)). They found essential genes to be significantly more evolutionary conserved than nonessential genes, in accordance with Jordan et al.'s findings. Focusing on *Pseudomonas aeruginosa*, (Dötsch et al., 2010) found a significant correlation between gene essentiality and evolutionary conservation, a correlation which, although weakened when gene expression was accounted for, remained significant.

These studies suggest that essentiality is indeed correlated with amino acid substitution rates in bacterial proteins but possibly not in eukaryotes. Yeast seems to be an intermediate case, where essentiality is not a determinant factor of protein evolutionary rate, but differential dispensability might be (Rocha & Danchin, 2004). Bacteria seem to harbor a small but significant correlation between evolutionary rate and dispensability which remains valid even after gene expression levels are controlled for, but the intricate interplay between evolutionary conservation, gene dispensability and gene expression is not yet sufficiently understood (Wolf, 2006).
It is highly important to note a central caveat permeating all the studies reviewed above – in all those, gene essentiality was determined on standard laboratory media in which these organisms are studied – however, the extent to which such media reflect the organisms' natural habitats is an open question, which becomes a critical one when studying phenomena that are strongly media dependent such as gene essentiality.

Throughout this work we term organisms whose essential genes are significantly more evolutionarily conserved than their nonessential genes – **Knockout Rate (KOR) organisms**. We term **non-KOR**, bacteria that are not KOR and we term **anti-KOR**, bacteria whose nonessential genes are significantly more evolutionarily conserved than their essential genes.

### 2.2 Genome-Scale Metabolic Modeling (GSMM)

Cellular metabolism is the set of biochemical reactions needed by biological cells to maintain life. These processes allow the cells to maintain their proper function, grow and respond to changes in the environment. Metabolism is often altered in disease, highlighting the importance of metabolic analysis in drug discovery and in understanding cell mechanisms and modes-of-action (Töpel, Hofestäd, Scheible, & Trefz, 2006). Furthermore, metabolic processes involve the production of industrially important nutrients, which explains the growing interest in metabolic biotechnological engineering applications (Durot, Bourguignon, & Schachter, 2009). Most of the chemical reactions within the cell are catalyzed by specific proteins called enzymes. These reactions typically convert several metabolites, called reactants (or substrates), into several other product metabolites. From the collection of these reactions, highly complex metabolic networks are formed, involving thousands of cross-talking reactions. The complexity of these networks poses challenges for the
field of metabolic modeling, requiring a systems-level approach (Price, Reed, & Palsson, 2004). One such approach is the Genome-Scale Metabolic Model (GSMM), a mathematical model of an organism's metabolic network used to computationally (*in silico*) assess phenotypic properties using a variety of computational methods (Feist & Palsson, 2008). The GSMM reconstruction pipeline and possible uses are depicted in Figure 1.

**Figure 1 - Formulation and use of GSMM as a four-step process:** Step 1, the GSMM formulation process is based on a variety of high-throughput datasets (omics data), including a comprehensive assessment of the literature (bibliomic data). Step 2, all of the data types are used to reconstruct a list of biochemical transformations that make up the reaction network, as well as its underlying genetic basis. Step 3, the data contained in the reconstruction can be formally represented in the form of matrices and logical statements that can be mathematically characterized by a variety of methods. Step 4, the model enables a broad spectrum of applications facilitated by the application of computational tools (Feist & Palsson, 2008).

The core of a GSMM is the reconstruction of the metabolic network in a mathematically abstract manner that lends itself to computation. The metabolic network is composed of a set of chemical reactions that can be represented as a set of chemical equations (Figure 2a-b). Embedded in these chemical equations is the information describing the reaction stoichiometry (Palsson, 2006), that is, the set of
coefficients that determines the balance between the different biochemical compounds that are consumed and produced by each reaction (Figure 2b). The stoichiometric information of the metabolic network can be represented in a matrix form, referred to as the stoichiometric matrix and denoted by \( S \) (Figure 2c). The matrix is organized such that every column corresponds to a reaction, every row corresponds to a biochemical compound (a metabolite) and the entries are the integer stoichiometric coefficients. Thus, each column describes the balance between the different compounds carried by the corresponding reaction. Similarly, each row describes all the reactions in which a certain compound participates and thus represents the interconnections between reactions in the network. Capturing both chemical and network information enables the stoichiometric matrix to transform the network flux vector (containing the flux rate of each reaction in the network) into a new vector containing the time derivatives of all the metabolites' concentrations (Figure 2d).
Simulating the inner workings of metabolic networks requires computational tools that engage GSMMs according to biological reasoning. Cells operate under constraints that govern their behavior and limit their range of possible functions (Price et al., 2004). Constraint-based modeling (CBM) imposes the set of governing constraints on the space of possible metabolic behaviors and allows filtering out behaviors that are not biologically feasible in a large-scale manner, thus enabling the prediction of various metabolic phenotypes such as growth rates, nutrient uptake rates, by-product secretions and gene essentiality (Price et al, 2004). CBM has been used for a variety of applications including comparative metabolic analyses over multiple organisms (Blank et al, 2005; Lee et al, 2009), drug discovery (Gordana et al, 2005), metabolic flux analysis (Rantanen et al, 2008), studies of network evolution (Fong et al, 2005) and metabolic engineering tasks (Pharkya et al, 2004). Flux Balance Analysis (FBA) is the most widely studied CBM method (Kauffman et al, 2003; Varma and Palsson, 1994) which searches for an optimal steady state solution that maximizes a certain objective function among all feasible steady state solutions. In microorganisms, FBA assumes that the organism strives to maximize its growth rate (or biomass production). To implement this, an artificial reaction that drains
biosynthetic precursors in an appropriate ratio is added to the model (Varma, Boesch, & Palsson, 1993).

GSMMs have been used towards five major ends: (1) contextualization of high-throughput data, (2) guidance of metabolic engineering, (3) directing hypothesis-driven discovery, (4) interrogation of multi-species relationships and (5) network property discovery (Figure 3; Oberhardt et al, 2009).
Figure 3 - Uses of metabolic GSMMs: The building and analysis of metabolic GSMMs are shown in the left panels, and the five categories of uses of metabolic GSMM as highlighted above are described in the right set of panels (Oberhardt, Palsson, & Papin, 2009).

Fifteen years ago, the GSMM of *Haemophilus influenza* became the first genome-scale metabolic reconstruction to be published (Edwards & Palsson, 1999). Since then, the field of genome-scale metabolic network analysis has expanded rapidly thanks to modern genomic sequencing technologies which enable rapid collection of annotated genomes and subsequently the rapid reconstruction of metabolic networks. To date, more than 50 curated reconstructions have been published (Duarte, Herrgård, & Palsson, 2004; Feist et al., 2007), spanning all three domains of life: Eukaryota, Bacteria, and Archaea. These encompass dozens of bacterial and yeast species, including various pathogens and industrially relevant organisms, the model plant Arabidopsis (Mintz-Oron et al., 2012) and mammalian metabolic networks including mouse (Sigurdsson, Jamshidi, Steingrimsson, Thiele, & Palsson, 2010) and human (Duarte et al., 2007). As efforts to develop new models failed to keep up with genome sequencing, Model SEED, a web-based resource for high-throughput generation, optimization and analysis of genome-scale metabolic models, was developed (Henry et al., 2010). The Model SEED integrates existing methods and introduces techniques to automate nearly every step of the model reconstruction process. This work resulted in the automated reconstruction of over 3000 bacterial genome-scale metabolic models.
2.3 Metabolic Models Used in Study

The systems-level approach applied in the utilization of genome-scale metabolic networks enables evaluation of a variety of features, including gene essentiality, while considering an organism's complete metabolic network and under different growth conditions (Oberhardt et al., 2009; Price et al., 2004). Though many human curated models for bacteria are now available (Baumler, Peplinski, Reed, Glasner, & Perna, 2011; Duarte et al., 2007, 2004; Feist et al., 2007; González-Domenech et al., 2012; Henry, Zinner, Cohoon, & Stevens, 2009; Kim, Kim, & Lee, 2010; Liao et al., 2011; Thomas et al., 2009; Y. Zhang et al., 2009) only a handful complied with the criteria defined as necessary for this research (Methods). As we were interested in leading a large scale study, we further made use of automatically generated models from the SEED project (Henry et al., 2010) finally compiling a group of 69 models covering 58 different bacteria.

2.4 Evolutionary Conservation of Essential Genes in Bacteria

This work is the first large-scale study of evolutionary conservation vs. essentiality in bacterial genomes. Using 69 metabolic models covering 58 bacterial species and integrating their evolutionary conservation rate estimates, we investigated whether bacteria adhere to the KOR hypothesis and how this tendency varies among different species. We present a novel method, Evolutionary Conservation Variation Analysis (ECVA), which calculates the minimal and maximal levels of evolutionary conservation of essential genes in an organism. Using the ECVA range, we place each bacterium along a spectrum of essential gene conservation from KOR to anti-KOR. Using this classification we tackle the question of whether bacteria tend to differentially conserve their essential genes and how this tendency varies among species.
To investigate whether bacteria with different tendencies for essential gene conservation have different phenotypes, we gathered a wide range of biological attributes (e.g. genome size, growth rate, habitat variety, phylogeny (Table_S2)) and computed the value of each attribute in each species (e.g. *E. coli* K12 has 4145 genes and *A. baumannii* is found in 11 environments, etc). We then looked for a significant separation in the values of these attributes between KOR/non-KOR bacteria and between different KOR classes. In the case of phylogeny, we searched for a phylum or class that was enriched with one type of bacterium. The same test was carried out for analysis of habitat variety, to look for a habitat enriched with one type of bacteria. A marked separation in any of the attributes tested between KOR and non-KOR bacteria would testify towards a biological manifestation of the tendency to conserve ones essential genes.
3 Methods

3.1 Biological Attributes

Table 1 includes a short description of the biological attributes used in this study. For the full descriptions, scores and references see (Table_S2 "Biological Attributes").

<table>
<thead>
<tr>
<th>Attribute Type</th>
<th>Reference</th>
<th>Score name</th>
<th>Scored models out of 69</th>
</tr>
</thead>
<tbody>
<tr>
<td>Environment (Organism Dependent)</td>
<td>Data for bacterial occurrence over different environments. Taken from Chaffron et al. (Chaffron, Rehrauer, Pernthaler, &amp; von Mering, 2010)</td>
<td>GG number of envs</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Data for bacterial occurrence over different environments. Taken from Chaffron et al. (Chaffron, Rehrauer, Pernthaler, &amp; von Mering, 2010)</td>
<td>GG environments</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Environmental complexity values obtained from Parter et al. (Parter, Kashtan, &amp; Alon, 2007). The natural environments of 117 bacterial species were ranked based on the NCBI classification for bacterial lifestyle <a href="http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi">http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi</a></td>
<td>6-class env complexity</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>Environmental complexity values obtained from Parter et al. (Parter, Kashtan, &amp; Alon, 2007). The natural environments of 117 bacterial species were ranked based on the NCBI classification for bacterial lifestyle <a href="http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi">http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi</a></td>
<td>6-class environments</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>Human Microbiome Project (Huttenhower C, Gevers D, Knight R, Abubucker S, 2012a, 2012b)</td>
<td>HMP sum (1,2,3)</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>Human Microbiome Project (Huttenhower C, Gevers D, Knight R, Abubucker S, 2012a, 2012b)</td>
<td>HMP unique(1,2,3)</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>Organism Dependent</td>
<td>Genome Size</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>Minimal duplication-time data were taken from (Couturier &amp; Rocha, 2006).</td>
<td>Duplication Time</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Taken from the ATGC database (Novichkov, Ratnere, Wolf, Koonin, &amp; Dubchak, 2009)</td>
<td>Phylogeny</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>For model sources see Methods</td>
<td>Model Complexity</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[AE]</td>
<td>69</td>
</tr>
<tr>
<td>Model Dependent</td>
<td>For model sources see Methods</td>
<td>AE percent</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[PE]</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PE percent</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>ECOEDS Dependent</td>
<td>EDC Growth</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>For the search algorithm (ECOEDS) see Results</td>
<td>Media Multiplicity Factor</td>
<td>69</td>
</tr>
</tbody>
</table>

Table 1 –Biological attributes used in the study: Biological attributes for the bacteria in this study. Bacteria were chosen for the study if they appeared in at least one of the biological databases and therefore, not all attributes are available for all bacteria (right column)
3.2 dN/dS Estimates

The basic measure of the selection pressure acting on protein-coding sequences is the ratio of the number of non-synonymous substitutions per non-synonymous site to the number of synonymous substitutions per synonymous site, also known as dN/dS. Substitutions are determined by comparing the genome in question with the genome of a closely related reference species or strain. Genes under purifying selection will display a lower dN/dS whereas positive selection, which increases dN, will result in increased dN/dS values (Novichkov, Wolf, Dubchak, & Koonin, 2009).

dN/dS estimates for some 400 bacteria were obtained from the Alignable Tight Genomic Clusters (ATGC) database (Novichkov, Ratnere, Wolf, Koonin, & Dubchak, 2009). Bacteria for the analysis were chosen from a set of 139 ATGC groups with the aim of having just one pair (bacterium and reference bacterium) per group. Also pairs were chosen so that the genome-wide synonymous substitution rate would be in the range of 0.25-1.5, preferentially choosing pairs such that the median dS will be as close to 0.75 as possible. The upper limit was chosen to keep dS below saturation (i.e., keep it a reliable estimate) for most of the genes in the genome, and the lower limit to reduce the fraction of proteins in the genome that have a very low nonsynonymous substitution rate (which are again unreliable) (Novichkov, Ratnere, et al., 2009).

3.3 Metabolic Model Processing

The following section details the different sources for metabolic models used in this study, as well as the processing done on them (Figure 4) and the criteria used in selecting them.
3.3.1 Model types

Two types of metabolic models were used: automatically generated and human curated. The automatically generated SEED models (Henry et al., 2010) were downloaded from http://seed-viewer.theseed.org/. A list of NCBI ID synonyms for the model genes were taken from the .fig files accompanying each model. This list was used to match between the model genes and the NCBI IDs from the ATGC dN/dS estimates (Methods). The human curated models were gathered from online resources; the full list of models and their reference article can be found in Table_S1 "Models DB". Some of the human curated models were given in SBML format and libSBML (Bornstein, Keating, Jouraku, & Hucka, 2008) was used to convert them to COBRA (Schellenberger et al., 2011) format. The match between gene names and NCBI IDs was either done by finding the curated model gene name (e.g. KPN_0002 for K. pneumoniae) in the corresponding SEED model .fig file (and then using the list of NCBI ID synonyms for this gene), or through .ptt files from the NCBI database ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/.

3.3.2 Separation of Exchange Reactions

Models with bidirectional exchange reactions pose a problem with regards to FBA flux solutions since a reaction can be consuming and secreting the same amount of a certain metabolite and show zero flux. Although this should not happen if the metabolite is used in any internal reaction, a transformation to separate intake and secrete reactions in the model simplified the code and model processing.

In order to separate the intake from the secretion in the flux solutions, each model's exchange reactions were duplicated into separate "intake" and "secretion" exchange reactions such that: For "intake", the S-matrix coefficients were positive, with bounds \(lb = 0, \ ub = \) previous \(lb\), limiting the reaction to intake only.
For "secrete", the S-matrix coefficients were negative, with bounds \( lb = 0, ub = \) previous \( ub \), limiting the reaction to secretion only.

### 3.3.3 Random Media

A set of random media was computed for each model. The random media were used as seeds for the Simulated Annealing (SA) search (detailed later on in Results) and as validation for comparison with the results of ECOEDS (Results). These media were sampled randomly from the metabolic model media space, with the only criterion for media validity being model viability (biomass production above some predefined threshold). To reduce the SA search space, availability of each compound in each medium was binary, i.e., the flux was either on or off. All bounds for reaction fluxes that were ‘on’ in a given medium were set to the same number (100), chosen to be high enough to surpass the (highest) positive lower bound some models place on certain constitutive reactions (such as ATP-production).

### 3.3.4 Essential Gene Calculation Preprocessing

The model-based prediction of essential genes for a metabolic model on a certain medium (Methods) was calculated many times during this study (\(~8M\) times for an SA search over all models). It was necessary to speed up this type of calculation as much as possible and so some preprocessing was done on all models to prepare data structures for quick calculations (Supp. Info).

### 3.3.5 Models Used In Study

In order for a model to be included in the study it needed to meet several criteria:

1. A Gene-Protein-Reaction (GPR) mapping must be present.
2. A biomass reaction must be present.
3. It must have non-zero maximum biomass under rich media.
4. We were able to map more than 50% of its genes to NCBI IDs.

5. Sufficient environmental information was present for the bacterium to allow the Biological Trait analysis (Results).

6. A reference organism was found in the ATGC database to produce a genome-wide synonymous substitution rate estimate in the range of 0.25-1.5.

All SEED models passed the first 3 criteria but surprisingly, there was a wide distribution regarding the percent of NCBI IDs that could be matched between the .fig files and the ATGC data (see Table_S1 "SEED without NCBI matches" for those that did not pass). Out of about 500 SEED models that passed criterion 5, only ~80 passed criterion 6. Out of the 40+ curated metabolic models gathered, ~25 passed the first 3 criteria and only 13 passed all 6 criteria. The final model list for this study includes 69 models consisting of 56 SEED models and 13 human curated models covering 58 bacterial species.
Figure 4 – Metabolic model preparation pipeline: The pipeline was applied to models acquired from SEED database (a) or from manually curated sources (b) which may need translation to Cobra format (c). SEED .fig files and NCBI .ptt files were used to map model genes into unique NCBI IDs from the ATGC database (d,e). Models exchange reactions are separated (f) and random media calculated (g). Models are preprocessed for quick essential gene calculations (h) and then used in ECOEDS SA search (i) and Purification (j), the result of which is a set of EDC media (k).
3.4 Calculating Essential Genes

A gene is considered essential under a certain medium \((\textit{in silico})\) if simulating its knockout (by constraining the flux through its GPR-associated reactions to zero) results in negligible biomass production. This process is repeated one by one for each gene. Given a metabolic model and a medium, the set of essential genes is uniquely defined and throughout this work, the terms ‘media’ and ‘essential gene sets’ may be used interchangeably (e.g., comparing media may correspond to comparing their essential gene sets). Since only viability and not growth rate is important, essentiality calculations are fairly robust to the definition of cellular objective function and using the biomass objective, the list of essential genes was found to be robust to threshold \((10^{-15} < \text{threshold} < 10^{-7})\).

3.5 AE, PE, APE and NE genes

For the split into "essential" and "nonessential" genes to be media independent, we defined the notion of \textit{Essential gene sets}: Using the metabolic models, genes that are found to be essential under rich media (all metabolite intake reactions are open) will be essential under any medium, since any medium is a proper subset of the rich medium. We call these \textit{Always Essential} (AE) genes. We term \textit{All Possible Essential} (APE) genes, those genes for which there exists a medium (even if only a single one) under which they are essential. Genes that are not in APE are termed \textit{Never Essential} (NE) and genes that are APE but not AE are termed \textit{Potentially Essential} (PE).

Clearly

\[
|APE| + |NE| = |AE| + |PE| + |NE| = |\text{model genes}|
\]

\[
AE \subseteq APE, \ AE \cup PE = APE
\]
Directly computing the APE set is difficult but a good approximation of it can be 
made. To approximate the APE set, the essential genes of several media types were 
calculated. Any gene found to be essential on any one of the media was included in 
the APE genes. The media used for the approximation:

1. EDC media found in the Evolutionary Conservation of Essentiality Directed 
   Search (ECOEDS) (Results)

2. A large set of random media (thousands of instances per model).

3. Media from an ECOEDS variant whose objective function was to find a 
   medium with the maximal number of essential genes.

A depiction of tests done to check whether the APE set gathered is complete enough 
can be found in the Supp. Info. APE can be directly calculated over any finite media 
as was done EDC media (EDC APE) and for Random Media (RD APE) (1,2 above).

A graphical representation of the Essential Gene Sets is shown in Figure 5 and a 
comparison between the set sizes across all models is shown in Figure 6 (Figure data 
found in Table_S1 "Essential Gene Sets").

Figure 5 –AE, APE and NE gene sets: A schematic representation of the Essential Gene 
Sets within a genome.
Figure 6 – Size of Essential Gene Sets across bacteria and media: AE is not media dependent and APE was estimated on all possible media. Random Media (RD APE) and Extreme Differential Conservation (EDC APE) were calculated on these media accordingly. The plot is arranged to display species sorted by their APE.

3.6 Definition of KOR and non-KOR Split

Using the definition of Essential Gene Sets we define several ways to split the metabolic models (and therefore their bacteria) into two groups, KOR and non-KOR for the sake of later analysis.

In two previous studies (Dötsch et al., 2010; Jordan et al., 2002) where bacterial essential genes were found to be more conserved than non-essential, researchers used experimentally determined gene essentiality of several bacteria, thus partitioning the genome into two groups – essential and nonessential. The dN/dS (or just dN) values for genes in these two groups were then compared. The problem with experimentally determined gene essentiality is twofold: First, experiments are time consuming and data is currently available only for a handful of species (R. Zhang & Lin, 2009).
Second, experiments performed on synthetic media may result in a different set of essential genes from those essential in a bacterium's natural environment where its genes have evolved (Rocha & Danchin, 2004). The first problem can be sidestepped by using *in silico* metabolic models, which allow us to assess gene essentiality on any bacterium with a GSMM and over a host of different media. However, metabolic model calculations of essential genes are still media dependent. We therefore defined three media independent methods to partition the genome into "essential" and "nonessential" using the Essential Gene Sets, as summarized in Table 2.

<table>
<thead>
<tr>
<th>Partition Name</th>
<th>Gene Group A</th>
<th>Gene Group B</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE</td>
<td>AE genes</td>
<td>All-but-AE genes</td>
<td>This is equivalent to the partition done in previous research had the experimental essential gene set been determined on a rich medium.</td>
</tr>
<tr>
<td>APE</td>
<td>APE genes</td>
<td>NE (All-but-APE) genes</td>
<td>This is similar to the partition done in previous research had the experimental essential gene set been determined on a poor medium.</td>
</tr>
<tr>
<td>AENE</td>
<td>AE genes</td>
<td>NE (All-but-APE) genes</td>
<td>This produces a more marked separation possibly helping overcome metabolic model inaccuracy. AE is the core group of essential genes with a higher probability of aligning with essential genes from experimental data. Similarly, NE is the group of genes that are never essential and are more likely to overlap with nonessential experimental genes. This partition does not cover all the (metabolic) genome, leaving out the PE set.</td>
</tr>
</tbody>
</table>

*Table 2 – Media independent essential gene partitions:* Three methods for partitioning the genome into “essential” and “nonessential” gene sets which are not media dependent. Methods rely on the definition of Essential Gene Sets.
Once a split is defined, dN/dS estimates of the two groups are compared to see if they significantly differ. Here too, we used two statistical tests to check for significance:

1. One-sided Wilcoxon rank-sum test assuming the median value for Group A (the essential group) is lower than for group B.

2. Single tailed t-test for log dN/dS, since bacterial dN/dS approximately follow a log-normal distribution (Novichkov, Wolf, et al., 2009).

Combined, this results in six methods (three splitting methods with two different statistical tests) assigning significance to the level of differential evolutionary conservation of essential vs. nonessential genes, as summarized in Table 3.

<table>
<thead>
<tr>
<th>Split Name</th>
<th>Description</th>
<th>Split Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE&lt;sub&gt;r&lt;/sub&gt;</td>
<td>AE split with rank-sum test</td>
<td>AE&lt;sub&gt;t&lt;/sub&gt;</td>
<td>AE split with t-test</td>
</tr>
<tr>
<td>APE&lt;sub&gt;r&lt;/sub&gt;</td>
<td>APE split with rank-sum test</td>
<td>APE&lt;sub&gt;t&lt;/sub&gt;</td>
<td>APE split with t-test</td>
</tr>
<tr>
<td>AEN&lt;sub&gt;E&lt;/sub&gt;&lt;sub&gt;r&lt;/sub&gt;</td>
<td>AENE split with rank-sum test</td>
<td>AEN&lt;sub&gt;E&lt;/sub&gt;&lt;sub&gt;t&lt;/sub&gt;</td>
<td>AENE split with t-test</td>
</tr>
</tbody>
</table>

**Table 3 –Methods for determining significance of differential conservation:** Three genome splitting methods along with two different statistical tests for significance lead to six methods to assign a significance-of-split score to a metabolic model genome.

For each of the 6 methods of assigning significance, a resulting p-value is obtained for every bacterial model; we call this the KOR score (Table_S1 "KOR Scores"). Model KOR scores from the 6 methods are highly correlated (average Spearman’s $\rho = 0.707, p < 0.0001$). Note that a different KOR score may be obtained for two (different) models of the same bacterium, for although the evolutionary conservation values are the same, the two models may incorporate a different set of metabolic genes and furthermore, their essential gene calculations may not always agree.

The KOR score allows us to rank the 69 models from the best (smallest p-value; a model whose essential genes are the most evolutionarily conserved compared to its
nonessential genes) to the worst (largest p-value; a model whose essential genes are the least evolutionary conserved compared to its nonessential genes) for a total of 6 different rankings. For each ranking of the models, a threshold is needed to separate them into two groups (KOR and non-KOR). We tested 6 different thresholds (10, 20, 30, 40, 50 and 60), each threshold determining the number of models in the KOR group. For example, for threshold=20, the 20 models with the best KOR score (lowest p-value) were put in the KOR group and the rest in the non-KOR group. Using 6 rankings (one for each method for assigning significance) and 6 thresholds for each we produce 36 ways to split the models into KOR and non-KOR bacteria. These will be referred to as the KOR vs. non-KOR splits and they will be used for analysis of biological attributes as will be seen.

3.7 Correcting for Multiple Hypotheses

In (section 4.5 Results) we used the 36 KOR vs. non-KOR splits for testing the separation of biological attribute values between the two groups. These tests are, however, highly dependent and so no correction for multiple hypotheses was done (and none was needed as there were none found significant even without correction). For the Hypergeometric tests on the phylogenetic tree (Results), each class and phyla were checked for enrichment. The bacteria in the study cover 7 phyla and 10 classes and so using the Bonferroni method, we deemed a phylum significantly enriched if its Hypergeometric p-value was below 0.01/7 and a class significantly enriched if its Hypergeometric p-value was below 0.01/10. The same kind of correction was used with the enrichment scores in 6-class and GreenGenes data (Results), each time accounting for the number of different environments.
Finally, each type of test – Separation between two groups, Hypergeometric enrichment in phylogeny and Hypergeometric enrichment in environments can be seen as a different (and somewhat independent) hypothesis. Therefore, a factor-of-three multiple hypotheses correction is due across all experiments. In most cases this was not used since even without it, no significant values were found. This additional factor is discussed in the phylogenetic tree Hypergeometric test results (Results).
4 Results

4.1 Overview

In this large-scale study of evolutionary conservation vs. gene essentiality, we examined three topics:

1) Do bacteria differentially conserve their essential genes?

2) How does the differential conservation of essential genes vary among bacteria and as a function of media?

3) Is the evolutionary conservation of essential genes in bacteria biologically meaningful?

To tackle the first two questions we present a new computational method, Evolutionary Conservation Variation Analysis (ECVA), which computes the range of differential conservation of an organism's essential genes over all possible media. Using ECVA we were able to position each bacterium along the KOR spectrum from KOR to anti-KOR. Bacteria were found to generally lean towards the KOR hypothesis, although species were found to occupy every position along the spectrum. We discerned a large variation in the tendency to conserve essential genes in bacteria and what's more, some model organisms used so far in studies of this sort were found to lie in the middle of the spectrum, such that they cannot be expected to show a disposition towards the KOR hypothesis, possibly explaining why so many contradicting results have been obtained to date.

Taking a different approach, we defined Essential Gene Sets that partition the genome into essential and nonessential genes in several media independent ways. We then computed these gene sets for the 69 models in the study and using evolutionary
rate estimates for the bacterial genomes we gave each bacterial model a KOR score, which is a measure of the evolutionary conservation of its essential genes relative to that of its nonessential genes. Here too, the distribution of KOR scores among bacteria strengthens the previous result that bacteria tend to differentially conserve their essential genes.

We tackled the third question by using the KOR vs. non-KOR split over a wide range of bacteria biological attributes (e.g. genome size, growth rate, habitat, phylogeny (Table_S2)), we searched for a significant separation in the values of these attributes between the two groups.

Surprisingly, no significant separation of biological attribute values was seen between KOR and non-KOR bacteria. That is, based on the comprehensive analysis done in this study, bacteria who comply with the KOR hypothesis have no biological phenotypes that distinguish them from those who do not. We also show that no habitat or phylogenetic group (with one possible exception) are enriched with either KOR or non-KOR bacteria, strengthening the previous conclusion.

4.2 ECV and ECOEDS

Here we present the Evolutionary Conservation of Essentiality Directed Search (ECOEDS) algorithm and the Evolutionary Conservation Variation Analysis (ECVA) methods. We will first give a brief description of both and then give an overview of ECOEDS and its stages.

ECOEDS is a two stage optimization algorithm that receives as inputs a GSMM, a random starting medium and evolutionary conservation scores for the model genes, and outputs a medium which maximizes the separation between the dN/dS values of
the essential and nonessential gene sets calculated on that medium. We term the resulting medium the Extreme Differential Conservation (EDC) medium. The first stage of ECOEDS uses Simulated Annealing (SA) to find a medium maximizing the KOR score within the media space and then, a Purification stage filtering redundant intake (exchange) reactions while yet lowering the KOR score.

We term the algorithm aiming to minimize the evolutionary conservation values of essential genes while maximizing those of the nonessential genes – ECOEDS (minimizing the KOR score), while we term the version that maximizes the evolutionary conservation values of essential genes while minimizing those of the nonessential genes – anti-ECOEDS (maximizing the KOR score). 250 ECOEDS iterations were performed on each model, each from a different random starting point and with a different random seed. From the resulting 250 EDC media the best scoring medium was chosen for the analysis.

ECVA was designed to overcome the obstacle of unknown bacterial natural media. Essentiality data obtained on synthetic media may not correspond to the set of genes essential in the bacterium's natural environment. Without knowing which are the 'true' essential genes, the real level of differential conservation of essential genes cannot be determined. ECVA overcomes this by searching for media that maximize and minimize the differential conservation of essential genes. Regardless of the natural medium’s true composition, the differential conservation of essential genes it induces lies within the bounds found by ECVA. ECVA was performed by running ECOEDS and anti-ECOEDS for each model and the two results, which offer the theoretical extremes in differential conservation of essential genes, enabling us to classify bacterial metabolic models into KOR classes as described in 4.3 (Results).
ECVA thus defines a fine grained classification of bacterial metabolic models along the evolutionary conservation of essential genes spectrum, which we call the KOR spectrum.

4.2.1 ECOEDS SA stage

The SA search stage was conducted over the metabolic-model's metabolite intake space. To simplify the search, media were reduced to a binary version, in which each metabolite was either present or absent. The media space can therefore be described as all possible binary vectors whose length is the number of intake reactions in the model (this differs from model to model, from 58 for *Buchnera aphidicola* to 341 in *Salmonella enterica*).

The SA search, running over the media space of a given metabolic model, requires a scoring function, which maps each point to a real number in order to converge to an optimal solution. To calculate this score for a given point (medium), the essential genes were calculated and then the dN/dS estimates for the genes were separated into two groups, corresponding to essential and nonessential genes as determined on the medium. The score was defined to be $\log(p)$ of the one-sided Wilcoxon rank-sum p-value, assuming the median of the essential group should be lower (more conserved) then the median of the nonessential group. By using the same score only substituting 'left' for 'right' tail, one obtains the anti-ECOEDS algorithm used in ECVA (Results). The SA search proceeded for a specified number of iterations and returned the best medium (lowest $\log(p)$) encountered (Supp. Info).

4.2.2 ECOEDS Purification Stage

The Purification stage filters redundant open intakes in the media output from the SA stage. The average resulting medium from the SA stage tended to have many open intakes, while solutions for optimal growth on such a broad medium tend to
utilize only a subset of these available intakes. To reduce a medium to include only necessary reactions (pure form), intakes were chosen at random and removed if viability was kept and the KOR score was the same or better (very small negative deviations from the score were permitted). This process resulted in a medium which had the same KOR score as that returned by the SA search (or even better) and with a far smaller set of open intake reactions. This stage represents another form of search towards the optimal medium with the best score, since the score can be significantly improved by this stage as well (Supp. Info).

4.2.3 ECOEDS efficacy

There are several reasons to believe that ECOEDS results in the absolute optimal score or very close to it: 1) ECOEDS is shown to outperform random media in 6.5 (Supp. Info) and synthetic media in 6.4 (Supp. Info). 2) Running ECOEDS 250 times for each model resulted in 250 EDC media which were never all unique. Counting the number of unique results and adding more runs (up to 2000) did not change the number of unique results. 3) The score of the unique media was very similar and one can assume that is a new cluster unique result were to appear, it's score would not significantly improve the global minimum if at all.

4.3 Evolutionary Conservation Variation Analysis (ECVA)

As described, ECVA is performed by running ECOEDS and anti-ECOEDS. For each model, both tests resulted in an EDC medium, called EDC and anti-EDC respectively. We computed the model essential genes on the EDC and anti-EDC media and for each, we calculated two KOR scores (one sided Wilcoxon rank-sum test): once with a 'left' tail (assuming the median dN/dS for the essential genes is less than the median for the nonessential genes) and once with a 'right' tail. We classified
as "KOR" those media (EDC or anti-EDC) who showed a significant result with the 'left' tail, "anti-KOR" those media who showed a significant result with the 'right' tail (media whose nonessential genes are significantly more evolutionarily conserved), and the rest were classified as "Undecided". We then classified all bacteria models into 5 KOR classes according to the classes obtained by their two EDC media. Table 4 summarizes the results and figure 7 depicts the classification scheme. For a full list of bacterial models and their KOR classes see (Table_S1 "KOR Classification").
Figure 7 – KOR classification scheme: Each metabolic model is processed twice, once with ECOEDS (a) and once with anti-ECOEDS (b). Both runs result in an EDC medium maximizing the differential conservation between essential and nonessential genes (c,d). After essential genes are computed on the medium, a Wilcoxon rank-sum test is run on the dN/dS of the essential and nonessential gene groups (e) resulting in a classification: if the result is not significant – the run in question is classified Undecided (f), if significant with essential genes more conserved – KOR (g) and if significant with nonessential genes more conserved – anti-KOR (h). The final classification into the 5 KOR classes depends on the combined results from both runs (i).
<table>
<thead>
<tr>
<th>Terminology</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>KOR Class Name</strong></td>
</tr>
<tr>
<td>Strongly KOR</td>
</tr>
<tr>
<td>Weakly KOR</td>
</tr>
<tr>
<td>Undecided</td>
</tr>
<tr>
<td>Weakly anti-KOR</td>
</tr>
<tr>
<td>Strongly anti-KOR</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>KOR classification</strong></td>
</tr>
<tr>
<td>Number of models</td>
</tr>
<tr>
<td>Representative Models</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

**Table 4 –KOR classification results:** Definitions of the KOR classes and the number of metabolic models in each class. Weakly KOR bacteria may be found to comply with the KOR hypothesis in some tests and not in others, depending on the media used for the experiments determining the essential gene sets. A 'c' before the bacterium name (e.g. "cS.oneidensis") signifies a manually curated metabolic model.

Note that the KOR classes are listed in order, from Strongly KOR (complies with the KOR hypothesis) to Weakly KOR (may or may not comply with the KOR hypothesis), and finally to Strongly anti-KOR (completely contradicts the KOR hypothesis). A comparison between KOR classes and the KOR vs. non-KOR split used in the analysis of the biological attributes is depicted in figure 8.

Several things should be noted:

1. No models were found to be strongly anti-KOR as opposed to 4 found to be strongly KOR. Approximating a normal distribution of the number of models in the 5 classes shows the mean to lie between Weakly KOR and Undecided. **This supports the claim that bacteria tend to follow the KOR hypothesis** (although weakly) as depicted in Figure 9 and Figure 10.
2. Several contradicting studies that aimed to determine if bacteria follow the KOR hypothesis examined essential genes in *E.coli* K12 ((Gong et al., 2008; Jordan et al., 2002; Rocha & Danchin, 2004). As seen in Table 4, we classified one model of this bacterium as Weakly KOR and the other as Undecided. This means that whether its essential genes are significantly more evolutionary conserved or not, is media dependent. Though essentiality data over the full genome may have presented a different picture, it is possible that this bacterium is a problematic candidate for testing the applicability of the KOR hypothesis to bacteria. The same KOR classification was done on the yeast *S. cerevisiae* metabolic model ([Supp Info](#)). It too was classified as Weakly KOR, which again may explain contradicting studies done with yeast.

3. 24 models were found to be Weakly KOR, which means they may follow the KOR hypothesis in some experiments and not follow it in others, highlighting that different datasets of essential genes may lead to contradictory results.

4. On certain media, some models were found to have significantly conserved nonessential genes compared to their essential genes, the complete opposite of the KOR hypothesis.
Figure 8 – KOR classes and KOR/non-KOR split comparison: Each error bar represents a model with the ECVA EDC and anti-EDC classes. KOR, Undecided and anti-KOR written in black are the EDC classes. The KOR vs. non-KOR split depicted with white on black background is threshold dependent. Several thresholds are shown and background is split according to threshold=10. (a,b) will be included in the KOR side whereas (c) will not. Similarly, for a threshold=60 split, (d) would be in the KOR side and (e) would not.
Figure 9 – Distribution of models along KOR spectrum: Each horizontal hourglass represents a model in the study with the ends depicting the KOR score of ECVA EDC media. The scores are log(p) significance of separation between dN/dS values among essential and nonessential genes. Both *E. coli* models used in the study are shown in brown (upper one is SEED model).
4.4 Are Bacterial Essential Genes more Evolutionary Conserved?

Aiming to detach ourselves from any specific medium under which the essential genes are tested, we proposed 6 media independent methods for assigning significance to the level of differential evolutionary conservation (Methods). Each one of these methods results in a per-bacterium KOR score denoting the separation of dN/dS estimates between essential and nonessential genes (Table_S1 "KOR scores"). Using this data, we may ask, how many bacteria have a significant separation relative to what would be expected by chance? If evolutionary conservation were equally distributed among essential and nonessential genes, one would expect on average between 0 and 1 models (0.69 from 69) to exhibit a p-value below 0.01. Table 5

Figure 10 – Number of models in KOR classes: The number of metabolic models in each KOR class. As visualized with the bounding curve, distribution tends towards the KOR class, showing that bacteria tend to conserve the sequence of their essential genes.
shows the number of bacteria with a significant KOR hypothesis propensity under each method for assigning significance.

<table>
<thead>
<tr>
<th>Significance Assigning Method</th>
<th>Number of bacteria under $p = 0.01$</th>
<th>Binomial p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AEt</td>
<td>14</td>
<td>8.90E-15</td>
</tr>
<tr>
<td>AEr</td>
<td>6</td>
<td>6.36E-05</td>
</tr>
<tr>
<td>APEt</td>
<td>17</td>
<td>3.54E-19</td>
</tr>
<tr>
<td>APEr</td>
<td>10</td>
<td>1.88E-09</td>
</tr>
<tr>
<td>AENEt</td>
<td>13</td>
<td>2.20E-13</td>
</tr>
<tr>
<td>AENEr</td>
<td>7</td>
<td>5.79E-06</td>
</tr>
</tbody>
</table>

**Table 5–KOR scores for different methods of assigning significance:** The 6 methods for assigning significance and the number of bacterial models showing a significant separation between evolutionary conservation values in essential and nonessential genes. The right column shows the Binomial p-values that such a number of conserved models will be obtained by chance.

As evident, many more bacteria have significantly lower evolutionary conservation values for their essential vs. nonessential genes than would be attributed to chance, strengthening the result of 4.3 (Results). This leads to the conclusion that overall, bacteria tend to conserve the sequence of their essential genes more than that of their nonessential genes.

Note that while ECVA assigned ranges according to essential genes calculated on specific media, this analysis used the notion of Essential Gene Sets that are media independent and do not correspond to essential genes over any specific media (except for the AE split which corresponds to a rich medium).

**4.5 KOR splits and Biological Attributes**

To check whether the evolutionary conservation of essential genes in bacteria relates to biological properties, we split the 69 bacterial models into two groups, KOR and non-KOR (Methods). These splits induce respective splits on the values of biological attributes of these bacteria. For example, splitting the models into two
groups, one can ask if the genome length is significantly longer (or shorter) in the KOR group.

We used three biological attributes which classify bacteria lifestyle:

1) Data mapping bacteria into 153 environments ("greengenes") from Chaffron et al. (Chaffron, Rehrauer, Pernthaler, & von Mering, 2010). Each bacterium receives a score consisting of the number of environments it was mapped to.

2) Data classifying bacteria into one of 6 lifestyles ("6-class") according to the multiplicity of environments it inhabits was taken from Parter et al. (Parter, Kashtan, & Alon, 2007). Each bacterium received a score from 1-6, 1 being obligate and 6 terrestrial (highly versatile) bacteria.

3) Bacterial samples from several locations on the human body ("HMP") was taken from the Human Microbiome Project (Huttenhower C, Gevers D, Knight R, Abubucker S, 2012).

We used three attributes that characterize bacterial metabolic models and genome:

1) Genome size - from the ATGC database (Novichkov, Ratnere, et al., 2009)

2) Model Complexity – defined as the number of non-zero elements in the metabolic model stoichiometric matrix.

3) |AE| - the size of the Always Essential gene group (Methods), computed on all metabolic models.

For more information on the attributes see (Table_S2 “Biological Attributes”).

For each attribute, we performed a statistical test over its values for KOR and non-KOR bacteria to see if they significantly differ. Checking with 36 different definitions of KOR and non-KOR bacteria over the 6 attributes, we found no significant difference between the two groups' values ($p > 0.01$ two sided Wilcoxon
rank-sum test). Figure 11 displays the mean±std for biological attribute values over the 6 methods for assigning significance (AEt, AEr, …, AENEr) and for a sample threshold of 20 (KOR group including the 20 best KOR scoring bacteria). Repeating the analysis on human curated models only, did not change the result and no significant difference between the biological attribute values of the two groups was seen ($p > 0.01$ two sided Wilcoxon rank-sum test).
Figure 11 – Biological attributes for KOR and non-KOR bacteria: 6 biological attributes over 6 significance assigning methods with 20 bacteria in the KOR group (results are robust to choice of threshold). Each plot shows the mean ± std of the two groups for each measure. The left (colored) bars represent values for KOR species and the right (blue) bars represent non-KOR species.

We performed a similar analysis by looking for a significant difference between the values of the biological attributes when split according to models in the 4 KOR classes (the Strong anti-KOR class being empty, was ignored Results). Using a-parametric ANOVA, no significant separation of values was observed on the previous 6 biological attributes (p > 0.01 Kruskal-Wallis test).
4.6 KOR Split and Phylogeny

To study whether the disposition towards the KOR hypothesis is phylogenetic, an effort was made to include models from a wide range of bacteria; Table 6 shows phyla and class data for the models in the study.

<table>
<thead>
<tr>
<th>Lineage Rank</th>
<th>Name</th>
<th># of models</th>
<th>Lineage Rank</th>
<th>Name</th>
<th># of models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phylum</td>
<td>Actinobacteria</td>
<td>3</td>
<td>class</td>
<td>Actinobacteria</td>
<td>3</td>
</tr>
<tr>
<td>Phylum</td>
<td>Clamydiae</td>
<td>1</td>
<td>class</td>
<td>Alphaproteobacteria</td>
<td>7</td>
</tr>
<tr>
<td>Phylum</td>
<td>Chlorobi</td>
<td>1</td>
<td>class</td>
<td>Bacilli</td>
<td>11</td>
</tr>
<tr>
<td>Phylum</td>
<td>Cyanobacteria</td>
<td>1</td>
<td>class</td>
<td>Betaproteobacteria</td>
<td>6</td>
</tr>
<tr>
<td>Phylum</td>
<td>Firmicutes</td>
<td>13</td>
<td>class</td>
<td>Chlorobia</td>
<td>1</td>
</tr>
<tr>
<td>Phylum</td>
<td>Proteobacteria</td>
<td>49</td>
<td>class</td>
<td>Clamydiia</td>
<td>1</td>
</tr>
<tr>
<td>Phylum</td>
<td>Tenericutes</td>
<td>1</td>
<td>class</td>
<td>Clostridia</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>class</td>
<td></td>
<td></td>
<td>Gammaproteobacteria</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>class</td>
<td></td>
<td></td>
<td>Mollicutes</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>class</td>
<td></td>
<td></td>
<td>Prochlorales</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6 –Phylogeny of bacteria in study: The phyla and class of all bacteria in the study are listed with the number of bacteria in each taxonomic group marked to the right.

For each of the methods for splitting bacteria into KOR and non-KOR, each phyla and class was checked for enrichment in either KOR or non-KOR bacteria. No such enrichment was found (Hypergeometric $p > 0.01$ corrected for multiple hypotheses (Methods)) apart from one exception: For the AEr split (Methods), with threshold=50, class Gammaproteobacteria was found to be enriched in non-KOR bacteria (Hypergeometric $p = 1.05 \times 10^{-4}$). After correction for multiple hypotheses the threshold for significance in this test is $0.01/10/3 = 3.33 \times 10^{-4}$, just slightly higher than the Hypergeometric enrichment score.

We performed a similar analysis by looking for significant enrichment of bacteria in one of the 4 KOR classes (Results). No such enrichment was found (Hypergeometric $p > 0.01$ corrected for multiple hypotheses).
As none of the splitting methods or KOR classes showed any significant bias towards any specific phylum or class, we suggest that the differential evolutionary conservation of essential genes is not a phylogenetic trait.

### 4.7 KOR Split and Habitat

To assess the connection between habitat and KOR bacteria, the 58 bacteria in the study were mapped to 77 different environments (Table_S1 "HG envs") using the GreenGenes databases (Chaffron et al., 2010). For each of the 36 methods for splitting bacteria into KOR and non-KOR, each environment was checked for enrichment in either KOR or non-KOR bacteria. To reduce noise, only environments with more than a certain number of applicable models (values of 10, 15, and 20 were tested) were used in the analysis. No enrichment was found (Hypergeometric \( p > 0.01 \) corrected for multiple hypotheses (Methods)).

Similarly, the bacteria were mapped to 6 lifestyles (Table_S1 "HG envs") by the 6-class database (Parter et al., 2007). As in the habitat analysis, none of the lifestyles were found to be enriched in either KOR or non-KOR bacteria or in any bacteria from the 4 KOR classes.
5 Discussion

In this study metabolic modeling was used on a large scale together with evolutionary conservation estimates to investigate three topics:

1) Do bacteria differentially conserve their essential genes?

2) How does the differential conservation of essential genes vary among bacteria and as a function of media?

3) Is the evolutionary conservation of essential genes in bacteria biologically meaningful?

The main novelty in our analysis was that it inspected the phenomenon of evolutionary conservation of essential genes in a manner which was not media dependent. This in turn allowed us to avoid the bias of essential gene sets captured on synthetic lab media. We used two different media-independent analyses: 1) splitting the genome into essential and nonessential genes, by means of the APE gene set which was estimated on all possible media and 2) ECVA analysis which estimates the range of differential evolutionary conservation.

We showed using both methods that weak evolutionary conservation of essential genes seems to be the norm for bacteria, since more cases of bacteria conserving the sequence of their essential genes vs. their nonessential genes were seen than would be attributed to chance alone (Results). We further demonstrated that bacteria lie on a wide spectrum of essential gene evolutionary conservation, from those who safeguard the sequence of their essential genes, to those who are indifferent, and on to those who seem to preserve the sequence of their nonessential genes (Table_S1 "KOR Scores", Results). Therefore,
depending on the bacterium at hand, different results may be obtained and contradictory conclusions drawn regarding the KOR hypothesis.

Looking for the source of the variability, we went on to test the association between the conservation of essential genes in bacteria and several biological phenotypes including genome length, growth rate, lifestyle, phylogeny and more. Splitting the bacteria to KOR and non-KOR, or according to the finer grained KOR classes, we were not able to find a biological attribute that unites any of the subgroups. Introducing bacterial habitat and phylogeny data, we again attempted to find a niche or evolutionary branch which is associated with bacteria in some KOR class, but did not find any significant results (save one instance of phylogeny), which suggests that the KOR trait is not generally linked to the phenotypic biological tendencies examined. Thus, the functional significance of the KOR hypothesis and the variance in evolutionary conservation of essential and nonessential genes remains unclear.

ECOEDS is an instance in a highly versatile family of algorithms looking for media with specific properties within the media space. In this work three variants were used, two maximizing the differential evolutionary conservation of essential vs. nonessential genes and one variant searching for media which produce the maximal number of essential genes (Methods). This was made possible by super quick calculation of the essential gene set (Supp. Info). In fact, any objective function calculated over metabolic models and involving the essential gene set can be run in this manner.
For example, a softer version of the KOR hypothesis may assume that in its natural environment, a bacterium’s essential genes are as conserved as possible relative to its nonessential genes. Using this weakened assumption, ECOEDS can be used to search for possible candidates to a bacterium’s natural media.

The larger evolutionary pressure expected on essential genes was thought to manifest itself in a slower evolutionary rate for these important genes (Hurst & Smith, 1999; Wilson et al., 1977). It may be that this is indeed what bacteria (and possibly higher Eukaryotes) "aspire to", but does not strongly manifest itself due to other forces at play that contradict or conceal this tendency.

Several possible shortcomings of this study and other factors that may have skewed the results are worth mentioning: 56 of the 69 models used were automatically generated by the Model Seed algorithm (Henry et al., 2010), having on average a 72% accuracy in their prediction of essential genes (Henry et al., 2010). As these are automatically generated models, we would expect their accuracy to be lower than that of the manually curated models. Still, limiting the study to the 13 human curated models did not change the final outcome. As this study was based on genome-scale metabolic models, only metabolic genes were taken into account. Across the full genome, AE, APE and NE would constitute different sets and the bacteria may have fared differently in the KOR vs. non-KOR split or in the KOR classification, leading to different results with regards to the separation of the biological attributes.

Bacteria and the metabolic models chosen for this study, although influenced by many independent variables, may present a bias and result in a set which does not properly represent the bacterial domain. With more than 3000 bacterial models in the
SEED database and dN/dS estimates for many of them on ATGC, using ECVA a much broader study encompassing a significant fraction of the bacterial species known is now possible and may be carried out.

Finally, the rate of synonymous to non-synonymous substitutions is measured by comparing the genomes of related species or strains, while essentiality is a trait of a single species. Gene importance may change during evolution, a certain gene having a backup gene in one species and not in another closely related one. Thus, some genes may be essential in species A but their ortholog may not be essential in species B, leading to a relatively large dN/dS for a truly essential gene, as measured in A (Gong et al., 2008).

This comprehensive systematic study presents both a novel metabolic model based algorithm – ECOEDS, and a new media-independent analysis method – ECVA. It is also of significance due to the large array of bacteria and attributes tested, far larger than any previous study dealing with evolutionary conservation of essential genes. It has brought new insights to an age old question regarding the evolutionary conservation of essential genes and has proposed possible explanations to the inconsistent results of past studies in this field.
6 Supplementary Information

6.1 Gene Essentiality Fast Calculations

Given a genome-scale metabolic reconstruction, the maximum biomass attainable under a specific medium can be computed using FBA and solving for maximum biomass. The effects of a gene knockout can be simulated by reducing the flux through all the reactions for which it is a necessary catalyst to zero and solving again for maximum biomass. Given that the maximum biomass under a certain medium is non-zero, a gene is called essential in-silico if the biomass computed under its KO constraints is zero (or numerically zero). Numerically zero is zero up to the numerical errors of the solver in use. In practice, a threshold of $10^{-12}$ was set as the upper limit for a numerical zero. Essential gene sets were shown to be robust with respect to this threshold up to several orders of magnitude.

For a certain GSMM, the set of essential genes under a certain medium are those for which the maximal biomass attainable under their KO constraints and using the medium intake constraints is numerically zero (Methods). In both the SA search, and the Purification stage, the mapping of a score to a medium necessitates the calculation of the set of essential genes for a certain media. For 70 models, running 250 SA searches each, with 200 search iterations, this sub function is called 3,500,000 times. In order to speed up the calculations, several aspects of the model were utilized as well as some optimization methods:

6.1.1 Static preprocessing per model

1. A gene is an essential catalyst of a reaction if it, or a complex of which it is a part, is the only one catalyzing the reaction. The GPR array was scanned and each gene $g$ was given a list of reactions for which it is an essential catalyst:
\[ Cat(g) = \{r_{j1}, r_{j2}, ..., r_{jn}\}. \]

2. Genes with an empty \( Cat \) list are never essential (NE) and were not considered.

3. A set of reactions \( R = \{r_{j1}, r_{j2}, ..., r_{jn}\} \) is said to cover a set of genes \( G = \{g_{i1}, g_{i2}, ..., g_{im}\} \) if

\[ \forall g \in G \quad Cat(g) \subseteq R \]

4. A set of reactions \( R \) is found, such that \( R \) is small compared to the full set of reactions in the model, but \( R \) covers as many genes as possible. This will be used by the dynamic processing later on per medium.

5. \( Cat(g) \) is not unique over all \( g \). Only unique sets of \( Cat \) were used in the calculations and the results were later mapped to all the genes.

6. A core set of AE genes were determined by calculating the essential genes on a rich medium (all intake reactions open). These genes will be essential in any medium and there is no need to recalculate their KO effect on biomass.

7. A further optimization is possible, by using the APE set (Methods) as a bounding set on all genes that may be found essential. Using this set, for a given medium, the search space for essential genes is greatly reduced, since only these genes need to be considered. This method is only approximate, since the exact set of APE genes is not known and hard to calculate directly. This method was tested but was eventually not used. It is mentioned here for the benefit of the reader which may choose to use it.
6.1.2 Dynamic processing per medium

8. An iMAT like Mixed Integer Linear Programming optimization (Shlomi, Cabili, Herrgård, Palsson, & Ruppin, 2008) is solved, looking for a reaction flux mapping to the network such that the biomass is non-zero and as many reactions in \( R \) carry zero flux.

9. If the MILP solution is successful, a set of reactions \( T \subseteq R \) is known to be able to carry zero flux when the biomass (or any other objective function) is non-zero. Genes covered by \( T \) are therefore not essential on this medium.

6.2 APE Completeness

The APE gene set for a certain model is complete, if it accounts for all the genes which could possibly be essential under any medium. The APE set can be said to be 'sufficiently complete', if encountering a medium which provides a new gene to the set is a rare event. To test this, for each model a large set (tens of thousands) of random media were computed and their essential genes calculated. No medium was found to add a new gene to the set (for each model) and so the APE sets in use were considered sufficiently complete.

6.3 ECOEDS SA and Purification Details

ECOEDS was run on Matlab R2013a using the CONDOR distributed parallel computation infrastructure. SA searches used Matlab's built in simulated annealing function (http://www.mathworks.com/help/gads/how-simulated-annealing-works.html) with a maximum of 200 iterations (adding iterations did not improve the best or average score of the final EDC media). SA searches were run with an exponentially decreasing temperature function: \( T(t + 1) = T(t) \cdot 0.95 \) (this function delivered the best results, i.e., the lowest p-value found).
We used a custom acceptance algorithm: given a current point with score $S_1$, a new point with score $S_2$ would always be accepted if $S_2 < S_1$ or if $S_2 \geq S_1$, with probability $\frac{1}{1 + e^{\frac{\Delta}{T}}}$ where $T$ is the current temperature and $\Delta = \frac{S_2 - S_1}{\text{Scale}}$ where $\text{Scale}$ was set to 0.05 being the average difference between p-values during the search.

We used a custom annealing algorithm: given a current point $p_1$ which is a binary vector of length $n$, the new point $p_2$ was determined as follows:

1. $R \leq \text{rand();}$
2. if ($R < \text{JumpThreshold}$)
3. $\text{Toggle} \leq \text{JumpSize;}$
4. else
5. $\text{Toggle} \leq \text{max(ceil((n/4) * (T/100)), ceil(n/100));}$
6. end
7. do until $p_2$ is viable
8. $p_2 \leq p_1$ with Toggle location flipped
9. end

Where $T$ is the current temperature, $\text{JumpThreshold}=0.01$ and $\text{JumpSize} = \lceil n/8 \rceil$. The algorithm flips a certain temperature dependent number of locations in $p_1$ while keeping the number not too large when $T$ is high and not too small when $T$ is low (line 5). It also performs a large jump from time to time (line 1-3). If the new point is not viable, a new point is generated. An environment was said to be viable if the result of an FBA run for biomass maximization was above some small threshold (Edwards, Ibarra, & Palsson, 2001; Varma et al., 1993). For each model, a small set of intake reactions was determined to be mandatory for biomass production (such as Zn, Mn, Mg and K intake reactions). These were kept open throughout the SA search since any media not containing them would not have been viable.
The purification stage constituted of several cycles, in each cycle intake reactions (which were open) were closed one by one, and remained closed if the medium without them was viable and induced an essential gene set which kept the current KOR score, or improved it. After all intake reactions were tested, another cycle would commence until a cycle where no intake reactions were successfully shut down, was reached.

A typical ECOEDS run on ~70 models with 250 different repetitions on each model with max 200 iterations in the SA search took around 18 hours. Although only 200 essential gene computations are needed in the SA stage, the Purification stage may require several hundreds more, depending on the number of cycles and the number of the open intakes. We estimate that over 8,000,000 whole metabolic model genome essential gene calculations were done on such a run.

Although starting from a random point in the media space, and the random nature of both the SA and the purification stages, for all models, of the final 250 media returned, the unique set of candidates was always lower than 250. This means that after purification, some searches converged to the same result. The attribute Media Multiplicity Factor relates to this (Table_S1 "Biological Scores").

### 6.4 EDC Media vs. Synthetic Media KOR Scores

ECOEDS and anti-ECOEDS search for media inducing the best and worst KOR scores respectfully. If the algorithm is effective, one would expect the KOR score of synthetic lab media to lie within the bounds of the EDC media KOR scores, i.e., the algorithm will find media which score better and worse than synthetic lab media. To evaluate this claim we used the Database of Essential Genes (DEG) to obtain experimentally ascertained sets of essential genes on synthetic lab media (R. Zhang & Lin, 2009). DEG covers 5 bacteria species that are included in our study, with 3
different experiments on *E.coli* K12. Table 7 summarizes the results. Using the experimental essentiality set as a reference, we gave the lab media a KOR score, i.e. the separation of dN/dS values induced by their split of genes into essential and non-essential sets. This score was compared to the best and worst KOR scoring medium found in ECOEDS for each bacterium. Across all bacteria included in the experimental set except one, then lab (experimental) media KOR scores lie within the bounds set by EDC Media (Binomial $p = 0.0029$). This also serves as a validation of the metabolic models' accuracy, since in almost all instances, the experimental dataset, when reduced to metabolic genes only, agrees with the results obtained from the metabolic models, i.e., the KOR score induced by the experimental set, lies within the bounds predicted by the metabolic models.

<table>
<thead>
<tr>
<th>DEG name</th>
<th>SEED exp. score</th>
<th>EDC score</th>
<th>Curated exp. score</th>
<th>EDC score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Most KOR</td>
<td>Least KOR</td>
<td>Most KOR</td>
</tr>
<tr>
<td><em>Burkholderia thailandensis</em> E264</td>
<td>-31.530</td>
<td>-24.069</td>
<td>-6.061</td>
<td></td>
</tr>
<tr>
<td><em>Escherichia coli</em> MG1655</td>
<td>-0.980</td>
<td>-1.245</td>
<td>-0.022</td>
<td>-4.096</td>
</tr>
<tr>
<td><em>Escherichia coli</em> MG1655 I</td>
<td>-0.854</td>
<td>-1.245</td>
<td>-0.022</td>
<td>-4.194</td>
</tr>
<tr>
<td><em>Escherichia coli</em> MG1655 II</td>
<td>-0.311</td>
<td>-1.245</td>
<td>-0.022</td>
<td>-2.447</td>
</tr>
<tr>
<td><em>Helicobacter pylori</em> 26695</td>
<td>-2.013</td>
<td>-2.145</td>
<td>-0.028</td>
<td>-5.004</td>
</tr>
<tr>
<td><em>Streptococcus pneumonia</em> R6</td>
<td>-0.095</td>
<td>-1.799</td>
<td>-0.050</td>
<td></td>
</tr>
<tr>
<td><em>Salmonella typhimurium</em> LT2</td>
<td>-0.503</td>
<td>-1.341</td>
<td>-0.028</td>
<td>-8.48</td>
</tr>
</tbody>
</table>

1.
Table 7 – Comparison of KOR scores obtained on EDC vs. synthetic media: KOR scores were computed on lab media (synthetic) according to experimental essential genes found on them. Values in red mark instances where lab media scored outside the bounds set by EDC media. Where available, both SEED and curated models were used for comparison.

6.5 EDC Media vs. Random Media KOR Scores

Similar to section 6.4, one would expect ECOEDS resulting EDC media to obtain a significantly better (lower) KOR score than random media. This was tested by calculating a set of Random Media (Methods) for each model and then calculating a KOR score for each medium in the set. For each model, the best KOR scoring random medium was compared to the best KOR scoring EDC medium as shown in Figure 12. EDC media always scored better (lower $\log(p)$) than random media, except for 2 cases, where the scores were even (for the R. prowazekii and U. urealyticum models). Since random media offer the starting points to the SA search, ECOEDS is expected to at least equal the KOR score of random media. The fact that 67 of 69 cases showed an improvement, testifies that ECOEDS offers a real optimization, consistently finding media with lower KOR scores.
Figure 12–Comparison of KOR scores between EDC and random media: The KOR score was computed on random media and for EDC media. Plot is arranged to display scores in increasing order.

6.6 KOR classification of S. cerevisiae

The metabolic model iMM904 (Mo, Palsson, & Herrgård, 2009) was used in the same manner as the other bacterial metabolic models. The KOR score ranges were from -23.230 to -1.3247 resulting in a Weakly KOR classification.

6.7 Models database and scores

Table_S1 includes all information about the models in the study and the biological scores:

- Sheet "Model DB" Holds the list of bacteria and their model attributes
- Sheet "Phylogeny" holds the phylogenetic info for the bacteria in this study.
- Sheet "KOR Scores" holds the KOR score for each model across the 10 different splitting methods.
- Sheet "Biological Attributes" includes all biological attributes used and their values for all models.
- Sheet "HG Envs" holds a list of the environmental habitats from two databases (Chaffron et al., 2010; Parter et al., 2007) and the number of models under each.

- Sheet "SEED without NCBI matches" includes a list of SEED models whose genes we were not able to pair up with enough NCBI IDs.

- Sheet "KOR classification" includes a list of all models and their corresponding KOR class.

Table_S2 holds a list of the biological attributes with information and references regarding each.


Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular systems biology, 3*, 121. doi:10.1038/msb4100155


