Title
Revealing the Genetic Basis of Natural Bacterial Phenotypic Divergence

Journal
Journal of Bacteriology, 196(4)

ISSN
0021-9193

Authors
Freddolino, Peter L
Goodarzi, Hani
Tavazoie, Saeed

Publication Date
2014-02-15

DOI
10.1128/jb.01039-13

Peer reviewed

JB
Journals.ASM.org

Revealing the Genetic Basis of Natural Bacterial Phenotypic Divergence

Peter L. Freddolino,[a] Hani Goodarzi,[a]* Saeed Tavazoie[a,b]

Department of Systems Biology, Columbia University, New York, New York, USA[a]; Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York, USA[b]

Divergent phenotypes for distantly related strains of bacteria, such as differing antibiotic resistances or organic solvent tolerances, are of keen interest both from an evolutionary perspective and for the engineering of novel microbial organisms and consortia in synthetic biology applications. A prerequisite for any practical application of this phenotypic diversity is knowledge of the genetic determinants for each trait of interest. Sequence divergence between strains is often so extensive as to make brute-force approaches to identifying the loci contributing to a given trait impractical. Here we describe a global linkage analysis approach, GLINT, for rapid discovery of the causal genetic variants underlying phenotypic divergence between distantly related strains of *Escherichia coli*. This general strategy will also be usable, with minor modifications, for revealing genotype-phenotype associations between naturally occurring strains of other bacterial species.

Despite the exponential increase in the number of sequenced bacterial genomes, enabled by next-generation sequencing, our ability to decipher causal genetic variants underlying phenotypic differences between strains has lagged far behind. Beyond enriching our basic understanding of bacterial evolution, this capacity is critical in many practical applications, for example, in pinpointing the genetic basis of differing intrinsic susceptibilities to antibiotics (1) or tolerances to organic solvents (2). Absent clear differences, such as the presence of a known antibiotic resistance mutation or gene, the presence of sequencing data alone generally does not allow identification of the genetic determinants of a given phenotypic variation, and the number of genetic differences even between closely related strains is often too large for a brute-force evaluation (that is, separate experiments in which each differing region is transferred from one background to the other) of all candidates to be practical. The development of more-efficient methods for identifying causal genotype-phenotype associations in natural bacterial strains is thus a crucial next step in taking advantage of the increasing wealth of available genomic sequences.

The challenge of identifying the genetic basis of naturally occurring phenotypic divergence is formally similar to the challenge of distinguishing between adaptive and hitchhiking mutations in laboratory evolution experiments (3–5) or between driver and passenger mutations in disease states (6, 7). A conceptually simple solution to this challenge is to swap alleles from the evolved to the parental strain, or *vice versa*, separately at all mutated sites. Given the number of mutations that are generally identified in laboratory evolution experiments, this approach, if deemed feasible, is very labor-intensive. We recently introduced ADAM (array-based discovery of adaptive mutations), a systematic experimental pipeline based on global linkage analysis, which employs a high-coverage transposon library to effectively perform numerous such allele swaps in parallel (8). ADAM has been applied successfully in several cases to identify the genetic variations underlying phenotypes in laboratory-evolved *Escherichia coli* strains (8, 9).

In principle, the same global linkage technology as that used in ADAM should enable the identification of causal genetic loci underlying phenotypic divergence between naturally occurring bacterial strains. However, the presence of nonhomologous regions will affect library representation and will have undefined and unexplored effects on the fitness scores obtained. In addition, whereas ADAM needs to detect changes only in a single direction (loss of the evolved phenotype), for comparisons between naturally occurring strains, fitness differences in either direction must be considered, and it must be possible to detect effects of various magnitudes. We have thus built upon the foundation of our previous approach to allow the use of global linkage analysis in distantly related bacterial strains, including the development of a new analytical framework which corrects for artifacts that render ADAM, in its original form, unsuitable in such cases. We refer to the combined experimental and computational methods for global linkage analysis of distant strains as GLINT (global linkage-based investigation of naturally occurring traits). Here we describe the GLINT method and showcase its ability to identify the genetic basis of natural phenotypic divergence at different levels of complexity, suggesting its general utility for rapid understanding of the genetic basis of phenotypic variation.

MATERIALS AND METHODS

Strains and media. All strains used in this study (see Table S2 in the supplemental material) were derived either from *E. coli* K-12 MG1655 (ATCC 700926) or from *E. coli* Crooks (ATCC 8739). We refer to the latter strain as ATCC 8739 throughout the text due to the presence of some ambiguity regarding its lineage (see http://ecoliwiki.net/colipedia/index.php/Talk:ATCC_8739); this strain was originally isolated from a human

Received 3 September 2013 Accepted 27 November 2013
Published ahead of print 6 December 2013
Address correspondence to Saeed Tavazoie, st2744@c2b2.columbia.edu.
* Present address: Hani Goodarzi, Laboratory of Systems Cancer Biology, Rockefeller University, New York, New York, USA.
P.L.F. and H.G. contributed equally to this article.
Supplemental material for this article may be found at http://dx.doi.org/10.1128/JB.01039-13.

doi:10.1128/JB.01039-13

fecal sample and has found use in a variety of laboratory tests (see, e.g., reference 10). Founder strains were obtained directly from the American Type Culture Collection. Routine growth and cloning were carried out in LB medium (10 g/liter tryptone, 5 g/liter yeast extract, 5 g/liter NaCl) or on LB plates (LB medium plus 15 g/liter agar). During cloning, the medium was supplemented with ampicillin (100 μg/ml), kanamycin (50 μg/ml), or chloramphenicol (30 μg/ml) (all from Sigma) as needed. All physiological experiments (except those explicitly noted as being performed in LB medium) instead used MOPS (morpholinepropanesulfonic acid) minimal medium (11) (Teknova, Inc.) supplemented with folate (4 μg/ml), with carbon sources and additives as described below. We use notation such as "MOPS-glucose-ellagic acid" to refer to MOPS minimal medium containing a carbon source (in this case, glucose) and an additive (in this case, ellagic acid); carbon sources were present at 0.2% (wt/vol) unless otherwise noted. 5-Keto-D-gluconate (KDG) was added from a 5% stock solution dissolved in water, with potassium hydroxide added as needed to ensure solubility. Ellagic acid (ELLA) stock solutions were prepared from a freshly opened vial of ellagic acid, dissolved to 1.25 mM in 500 mM NaOH, diluted in 500 mM NaOH to a working concentration of 250 μM prior to addition to culture media, and then used at a final concentration of 10 μM during all follow-up experiments (to minimize precipitation, which we observed in the presence of bacterial growth at higher concentrations). Ellagic acid-containing media proved particularly vulnerable to oxidation, and thus, the stock solution used in follow-up experiments was stored in frozen single-use aliquots, and the ellagic acid medium was prepared fresh immediately prior to use. For the library selections, we used a 5-fold-higher concentration of ellagic acid (final concentration, 50 μM; obtained from dilution of a 1.25 mM stock); the use of a lower concentration in follow-up experiments was to avoid precipitate formation, which we observed at 50 μM ellagic acid in the presence (but not in the absence) of bacterial growth. Due to the large amount of NaOH in the finished medium, the pH of the ellagic acid-containing medium immediately after preparation was approximately 9 (50 μM ellagic acid) or 9.5 (10 μM ellagic acid), and hence we consider the MOPS-glucose-ELLA stress condition to represent a combination of high pH and the presence of the ellagic acid itself.

All PCR primers used for strain construction and validation are shown in Table S3 in the supplemental material. Aside from the parental strains, all strains were validated by sizing of the P332/P333 product to confirm the MG1655 or ATCC 8739 origin and by sequencing of the location of at least some portion of the transferred region or scar expected to be present (Sanger sequencing by Genewiz, Inc.).

Condition screening. In order to identify growth conditions under which MG1655 and ATCC 8739 differ significantly in fitness, we obtained growth curves for both strains on Biolog Phenotype MicroArrays 1, 2, 3B, and 4A, which test for respiratory activity in a variety of nutrient sources (~400 conditions). In addition, we assessed the growth of both strains in the presence of all compounds from the DTP natural product set (obtained from the Developmental Therapeutics Program of the National Cancer Institute) in M9-glucose minimal medium with the compounds at 10 μM. All screens were performed at least in duplicate, and we identified all cases showing consistent differences between MG1655 and ATCC 8739 in growth rate, lag time, or the integral of the growth curve. We identified a total of 14 conditions (11 nutrient sources and 3 antibiotic compounds) showing consistent differences between strains, of which 3 were chosen for follow-up experiments.

Comparison of MG1655 and ATCC 8739 genomes. The alignment used in Table S1 and Fig. S1 in the supplemental material was obtained using Mauve, version 2.3.1 (12), with the progressiveMauve aligner and default parameters. The homologous blocks written by Mauve to an XMFA output file were analyzed to determine the numbers of direct matches, mismatches, and indels between the strains; any portion of either genome not present in those blocks was assumed to be unique to the corresponding strain.

Transposon library construction and transposon footprinting. The primary transposon library in the MG1655 background was prepared using previously published methods (13, 14). The secondary library in ATCC 8739 cells was prepared by P1*vir* transduction (15) and selection in kanamycin-containing media. Detailed protocols for each step are given in Text S1 in the supplemental material.

Construction of validation strains. All strains containing gene deletions or regions transferred from MG1655 to ATCC 8739 were generated using the system developed by Cherepanov and Wackernagel (16). For any particular insertion, a kanamycin resistance cassette was amplified from pKD4 using primers that yielded 40-bp regions of homology to the target site in the genome on either end. The targeting extensions were immediately adjacent to each other when only a tag was desired or flanked the region to be excised if tagged deletion was desired. Target cells containing pKD46 (17) were grown to mid-log phase in salt-free LB medium supplemented with ampicillin and arabinose (100 μg/ml and 10 mM, respectively), then electroporated with approximately 100 ng of the amplified resistance cassette, recovered for 1 h in SOC medium (18), and finally plated onto LB plates containing kanamycin. All knockouts and cassette insertions were initially performed in our MG1655 strain; if needed, the marked insertions were transferred to ATCC 8739 using standard P1*vir* transduction (18). Resistance cassettes were then excised by transforming marked strains with pCP20 (16), selecting on LB medium plus ampicillin at 30°C, and then growing transformants for 24 h on LB medium at 42°C. Colonies were then replica plated on LB medium, LB medium plus ampicillin, and LB medium plus kanamycin to ensure the loss of pCP20 and the resistance cassette. In all cases where a homologous region was transferred from MG1655 to ATCC 8739, we constructed strains directly in the ATCC 8739 background containing the same resistance cassette scar as that present in the region transferred from MG1655; unless otherwise noted, the resistance scar alone had no significant effect under the conditions that we considered (that is, 95% confidence intervals for the ratio of growth rates contained 1, and in the ellagic acid case, 95% confidence intervals for the difference in lag times contained zero). In the case of the *htrL* Kan cassette insertion used to transfer the *rph* locus, the cassette was in a region unique to MG1655, so we instead confirmed that the scar itself had no effect in the MG1655 background.

Selection of libraries under test and reference conditions. For each selective condition (see Table 1), the secondary library described above was thawed on ice, pelleted by centrifugation for 3 min at 17,900 × *g*, resuspended in an equal volume of carbon source-free MOPS medium, and then diluted 500-fold into 25 ml of prewarmed, preaerated selection medium. Libraries were incubated for 16 h at 37°C with shaking at 250 rpm and were then placed on ice. The selected libraries were harvested by centrifugation for 10 min at 5,525 × *g* and 4°C, and then the pellets were immediately lysed and processed for transposon footprinting as described below.

Transposon footprinting and microarray analysis. The abundances of different MG1655-specific insertions in the secondary library were measured by using transposon footprinting (14) followed by hybridizing samples from selective and reference conditions to an Agilent tiling microarray. Details of the transposon footprinting and hybridization procedures are given in Text S1 in the supplemental material.

Microarray spot intensities were extracted using Agilent Feature Extraction, version 9.5. Any spots for which the IsSaturated or IsFeatNonUnifOL flag was true, or for which the IsPosAndSignif or IsFound flag was false, were discarded. Log ratios of selected to reference signals for the probe at each position were estimated by pooling data across all biological replicates and replicate spots on the array, using the error-weighted combination method in section 3.1 of reference 19. The processed log ratios were then analyzed using the GLINT postprocessing pipeline as described below.

Postprocessing of insertion abundance data. Due to the unique characteristics of transposon insertion abundance/global linkage analysis (compared with, e.g., expression analysis of chromatin immunoprecipi-

tation data), we developed a novel postprocessing pipeline for normalizing global linkage analysis data and identifying regions that contribute significantly to fitness differences between strains. Here we describe the resulting algorithm; programs implementing this pipeline for the postprocessing of GLINT data are available for download from https://tavazoielab.c2b2.columbia.edu/GLINT/. Note that in the description of the algorithm here, we provide default values for several parameters (for example, the smoothing window width) that were used in our analysis but may be easily changed if necessary; see the documentation included with the program for details.

The probe-level log ratio abundance data (obtained in this case by two-color microarrays) were first normalized to have a median of zero and were scaled to have an interquartile range corresponding to that of a standard normal distribution (as in the analysis of reference 20). Next, the data were normalized for homology-dependent effects in two steps (see Results for the exploratory data analysis and justification leading to this procedure). All genomic regions in the donor genome that aligned to a gap greater than 2 bases in the recipient genome were removed and were replaced with a single representative probe at the center of the removed region, bearing as a selection score the median for all the probes that it replaced. The broadened distribution of selection scores in probes adjacent to large insertions was corrected by performing loess smoothing (21) of the magnitude of the selection score as a function of the distance from the nearest large insertion (here we used a value of 0.3 for the loess smoothing parameter) and then dividing the normalized log ratio at each point by the value predicted in the smoothed model; we henceforth refer to the resulting value as the selection score. The selection scores were then smoothed using a rolling median over a 21-probe window (approximately 1 kb on our tiling array) to provide the smoothed selection scores.

In order to assess the significance of the smoothed selection scores, we modeled the background distribution of insignificant probes as arising from an autoregressive model, in order to account for the substantial autocorrelation between adjacent probes (a hazard originally noted for experiments with tiling microarrays in reference 22, although we treat the problem in a somewhat different way). The mean and standard deviation of the modeled background were estimated to be the median of all normalized probe scores and the median absolute deviation of all normalized probe scores divided by the 75th percentile of a standard normal distribution, respectively. The correlation structure was estimated by averaging the empirical autocorrelations for all continuous windows of 41 probes or more, excluding probes with undefined values or those in the top or bottom 2% of all normalized scores. We then parameterized an autoregressive model (in this case, 23rd order) to match the observed autocorrelation function, with the mean and standard deviation obtained above. The order of the model was determined by examining the autocorrelation function of probes from the background distribution (excluding the high- and low-value probes as described above), which reached zero within 20 to 30 probes under each condition. The null distribution of the smoothed selection score was approximated by drawing 500,000 simulations of 21 probe windows from this distribution and calculating the median from each draw.

With the background distribution thus established, two-tailed *P* values were calculated for the smoothed selection scores at all probes, and significant probes were identified using the Benjamini-Hochberg procedure (24) with a false discovery rate of 1%. Any run of 10 or more significant probes with scores of the same sign was flagged as a prospective peak. Finally, to facilitate interpretation, prospective peaks with scores of the same sign separated by no more than 5 kb were lumped into peaks, and only lumped peaks of at least 5 kb in total length were retained (thus requiring either a single prospective peak of at least 5 kb or two or more prospective peaks covering a region of at least 5 kb with no large gaps between them); we report lumped peaks passing this final length filter as peaks throughout the paper. The lumping step applied here is justified by the biology underlying the construction of the secondary library; P1*vir* transduction transfers genomic regions tens of kilobases in length (18), so any true signal from a locus that has substantial fitness effects upon transfer should be sustained over a relatively large genomic distance. Tuning of the precise parameters used in the lumping and final peak size filter may be required for libraries constructed using other methods.

The output from this postprocessing pipeline is a set of peak calls indicating regions of the genome for which transfers from the donor to the recipient strain showed significant enrichments or depletions in the selected secondary library, alongside diagnostic plots showing the distribution of scores used in normalization steps and (optionally) the probe-level scores used at various points in the algorithm. The procedure used here is very similar to well-established methods for the analysis of ChIP-chip (chromatin immunoprecipitation with microarray technology) data (see, e.g., references 22 and 25), but extended to appropriate-length scales and with additional corrections dictated by the nature of the data used here, and particularly, extensions to correct for the presence of low-homology regions unique to GLINT. We anticipate that the same method would be applicable to data from applications of global linkage analysis using other methods for library construction or population abundance readout; possibly the distance and cutoff parameters used would need to be tuned in order to ensure that they are appropriate for the case at hand. In particular, the peak-calling module is usable in isolation on data from ADAM experiments, in which case all homology information is ignored.

While the method presented above is described in terms of analyzing microarray data, the same procedure could be used with minor preprocessing on data in which insertion abundances were instead measured using high-throughput sequencing. In our experience, it is convenient and computationally efficient for similar data sets (in which base pair-level resolution is not meaningful due to the long genomic distances over which the phenomena of interest occur) to represent the sequencing data as log ratios of properly normalized read densities under the selective versus unselective conditions, downsampled to representative points every 5 to 50 bp. Sequencing data prepared in this way will have properties very similar to microarray data (as long as counts are high enough to approximate the read densities as continuous), and the procedure outlined above should yield reasonable peak calls.

Growth curves. All growth curves were measured in Costar 96-well plates on either a Synergy MX or a PowerWave XS2 plate reader (BioTek, Winooski, VT). Each well contained 150 μl medium and 100 μl mineral oil (catalog no. BP26291; Fisher); both our experience and that of others (26) indicates that this use of mineral oil does not alter culture aeration, at least during the stages of growth during which lag times and growth rates were measured. Plates were incubated at 37 C with shaking ("fast" for the Synergy; "medium" for the PowerWave), and measurements of the optical density at 600 nm (OD_{600}) were taken once every 10 min for all wells. Cells were pregrown overnight in MOPS plus 0.04% glucose and were then diluted approximately 400-fold into prewarmed medium appropriate to the experiment. All minimal media used in plate reader experiments were supplemented with 0.001% Tween 20 (Sigma) to prevent aggregation-based artifacts. At least two biological replicates were obtained on different days for each strain-medium combination; each biological replicate comprised 6 to 11 independent technical replicates (wells) within a 96-well plate.

For growth on MOPS-glucose-ellagic acid medium, both lag times and growth rates proved to be important, and thus, modified protocols were used both for the experimental setup and for data analysis. Because lag times depend on the physiological state of the inoculum, for our purposes we defined the lag time as arising from cells growing in mid-log phase in MOPS-glucose medium. To this end, we pregrew cells for 5 h in MOPS-glucose medium immediately prior to transfer to a medium containing ellagic acid. Cells were pregrown in lidded 96-well plates, from a 400-fold dilution of an overnight culture in MOPS plus 0.04% glucose. Cells were then transferred as rapidly as possible to prewarmed MOPS-glucose-ellagic acid medium in a separate plate and were covered in prewarmed mineral oil, and growth was measured as described above.

The raw optical density data were processed using in-house scripts to

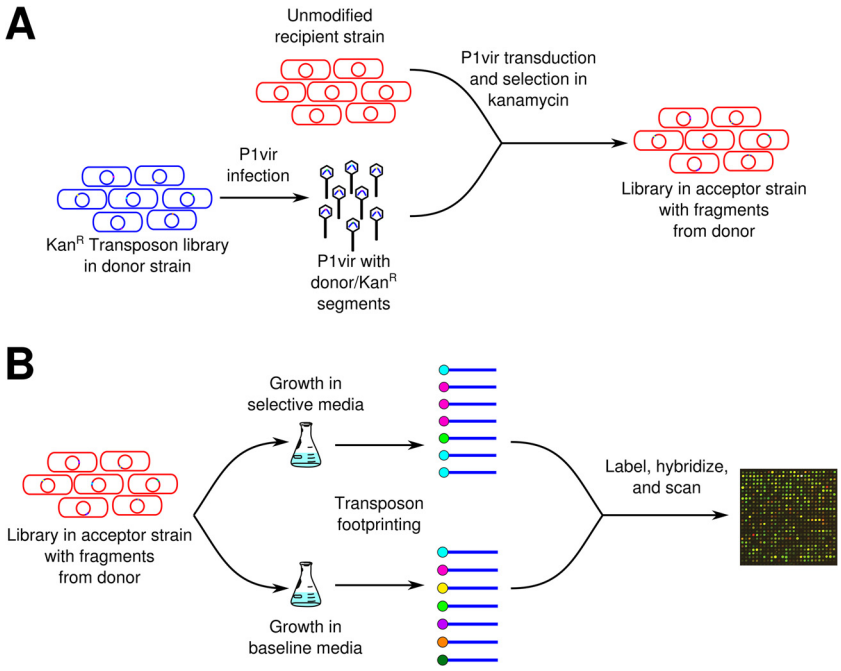

FIG 1 Schematic of the library generation and library selection steps required for cross-strain global linkage analysis. (A) Library generation. A library of cells in the donor strain is prepared with random integration of selectable markers at points throughout the genome (in this case, through random integration of a transposon containing a kanamycin resistance gene). This "primary library" is then used to prepare a "secondary library" in the recipient strain by random homologous recombination of genomic regions from the primary library into the recipient (in this case, using P1 transduction). The secondary library is limited by use of the selective marker, such that it contains only cells that received a marked genomic region from the donor strain. In both panels, genomic materials originating from the donor and recipient strains are shown in blue and red, respectively. In the application presented here, the donor and recipient strains are MG1655 and ATCC 8739, respectively. (B) Library selection. A pool of cells from the secondary library is grown in parallel under the condition of interest and a permissive reference condition. The abundances of DNA from the donor strain at sites throughout the genome are compared between the two conditions by quantifying the frequency of the resistance marker (in this case, by transposon footprinting and microarray analysis), which must have originally been transferred from the primary library.

determine the maximum growth rate and lag time for each replicate, and a linear mixed-effects model was then used to estimate the underlying parameters for each strain. Details of the analysis are given in Text S1 in the supplemental material.

Microarray data accession number. Raw microarray data have been deposited in the Gene Expression Omnibus as accession no. GSE45421.

RESULTS

The experimental component of the GLINT method is schematized in Fig. 1 (for details, see Materials and Methods). In brief, genomic regions surrounding transposon insertions are transferred from a donor strain to a recipient, yielding a library of cells with different portions of the donor genome in the recipient background (Fig. 1A). Genetic footprinting of the insertion sites using previously described methods (13, 14, 27) can then be performed on populations that are subjected to selective conditions (where the differential phenotype of interest will be displayed) alongside unselective conditions as controls (Fig. 1B). Comparison of transposon insertion abundances throughout the genome in these selected libraries, followed by postprocessing with the GLINT analysis pipeline, allows the identification of the primary contributors to fitness differences between the donor and reference strains under the selective condition of interest. As detailed below, few details of the experimental procedure are set in stone; the methods for tagging the donor genome, transferring that genomic material to the recipient strain, and measuring transposon abundances in the selected populations can all be varied so long as a few basic requirements are met. In the test application described below, the experimental procedures closely mirrored ADAM (8) except for the use of increased numbers of transductions that were subsequently pooled to yield a sufficiently diverse secondary library.

Regardless of the precise methods used for library generation, the application of global linkage analysis to distantly related organisms requires consideration of the effects of low-homology regions and large insertions or deletions on the resulting selection scores. As described in more detail under "Effects of homology on selection scores" below, correction of selection scores near large insertions or deletions, particularly near large insertions in the donor strain relative to the recipient, is necessary for those scores to accurately reflect the fitness effects of genetic differences at various points in the genome. We have incorporated an appropriate set of corrections into a postprocessing tool for GLINT data (freely available from https://tavazoielab.c2b2.columbia.edu/GLINT/) that is designed to automatically identify genomic regions causing substantial fitness differences between the strains under comparison, while making minimal assumptions about the experimental details used to generate selection score data. The stages of this postprocessing pipeline are schematized in Fig. 2. First, a global alignment of the donor and recipient strain genomes is used to identify donor-specific regions. Donor-specific regions are removed and replaced by a single representative value to prevent overrepresentation of these regions in the selection score results. In addition, since we have observed that the magnitudes of selec-

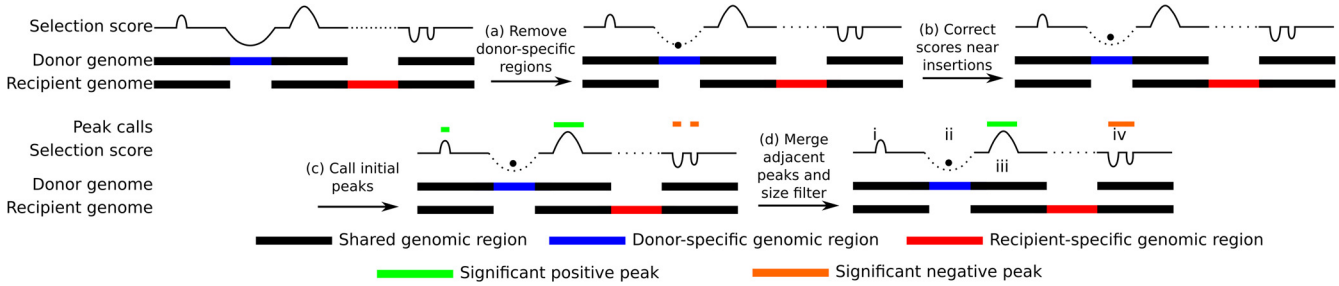

FIG 2 Schematic of computational postprocessing of GLINT experiments. As detailed in the text, donor-specific regions are each condensed into a representative point (a), and then scores near the donor-specific insertions are rescaled to bring their magnitude in line with the remainder of the genome (b). Candidate peak locations are identified by the presence of a run of significantly enriched or depleted selection scores (c). Finally, peaks are filtered based on their sizes; only regions with a high enough density of candidate peaks are included in the final peak calls (d). Thus, even regions with apparent peaks may be eliminated either because they are too narrow (i) or because they occur primarily due to the presence of a donor-specific insertion (ii). On the other hand, true peaks will be called either for a single large significant region (iii) or for a region containing several small, closely spaced candidate peaks of the same sign (iv).

tion scores are artificially inflated near donor-specific insertions, all selection scores are normalized by using a nonparametric curve fit to estimate the dependence of the selection score on the distance from the nearest large insertion. Significant peaks or troughs (we generally refer to both as "peaks") in the normalized selection score profile are then identified by assuming an autoregressive background model (which accounts for the probe-probe correlations inherent to tiling arrays), and then nearby peaks are merged and filtered based on their sizes. The default parameters used in the computational pipeline have been optimized for our specific experimental procedure; we provide guidance for determining appropriate values in Text S2 in the supplemental material.

As a test application of global linkage analysis to highly diverged strains, we phenotypically and genetically compared *E. coli* K-12 MG1655 (ATCC 700926; referred to below as MG1655) with *E. coli* strain Crooks (ATCC 8739); according to recent phylogenetic analysis, these strains belong to the most divergent subgroups present among the commensal phylogroup A *E. coli* strains (28). Comparisons of the MG1655 and ATCC 8739 genomes appear in Fig. S1 and Table S1 in the supplemental material, showing the presence of one large-scale genomic rearrangement and hundreds of smaller insertions or deletions between the two strains. In all, 7.5% of the MG1655 genome and 9.6% of the ATCC 8739 genome consist of regions not present in the other strain, whereas within alignable regions, the sequence identity is, on average, above 99%.

To verify our ability to characterize the genetic basis for physiologically relevant phenotypes between divergent *E. coli* strains, we performed a medium-throughput screen to identify conditions showing qualitatively apparent differences in growth between MG1655 and ATCC 8739. From a set of 14 candidates (see Materials and Methods for details), we chose 3 for further analysis, summarized in Table 1. For each of these three conditions, we used GLINT (with an MG1655 transposon insertion library as the donor strain) to identify the most significant loci contributing to fitness differences under that condition. Summary plots of the results for all three cases are shown in Fig. 3 and are tabulated in Table 2; in each case, only a small set of loci with significant selection scores is shown. Notably, different levels of complexity are apparent in the test cases, ranging from two to seven significant peaks. To validate the sites identified, for each condition, we chose one or more of those with the strongest selection scores and transferred them from MG1655 to ATCC 8739; as detailed below, all show fitness effects (measured by changes in growth rate or lag time) consistent with the linkage analysis results.

ATCC 8739 cannot grow on 5-keto-D-gluconate as a sole carbon source due to the lack of a single enzymatic functionality. As shown in Fig. 4B, ATCC 8739 grows extremely poorly on 5-keto-D-gluconate (KDG) as a sole carbon source, whereas MG1655 grows readily. The GLINT fitness profile using growth on KDG as a selective condition shows a single peak, centered around *yjhP*, immediately adjacent to the *idnDOTR* operon (Fig. 4A). D-Gluconate metabolism in K-12 strains may occur through the action either of the GntI system (*gntT*, *gntU*, *gntK*) or of the GntII system (*idnT*, *idnK*) (29, 30); however, KDG must first be reduced to D-gluconate by the product of *idnO* (30). Alignment of the MG1655 and ATCC 8739 genomes with Mauve (12) shows that *idnDOTR* is located in a K-12-specific portion of the genome, whereas all three GntI genes are in conserved regions. Furthermore, Megablast (31) searches for regions homologous to each of the *idn* genes in ATCC 8739 show that this strain contains no region with significant homology to *idnO* or *idnK*. Consistently,

TABLE 1 Summary of conditions considered here under which MG1655 and ATCC 8739 cells show substantial differences in growth

Comparison[a]	Selective condition	Key statistic	Fitted value (95% confidence interval) for key statistic	
			MG1655	ATCC 8739
KDG vs GLU	MOPS–5-keto-D-gluconate	Growth rate (doublings/h) under selective condition	0.391 (0.377 to 0.408)	0.0 (N/A)[b]
LB vs GLU	LB medium	Growth rate (doublings/h) under reference condition	1.030 (1.010 to 1.050)	1.365 (1.337 to 1.391)
ELLA vs GLU	MOPS-glucose-ellagic acid	Relative lag time (h) under selective condition	3.953 (2.895 to 5.042)	0.0 (−1.031 to 0.935)

[a] KDG, MOPS–5-keto-D-gluconate; GLU, MOPS-glucose; ELLA, MOPS-glucose-ellagic acid. The reference condition was MOPS-glucose in all cases.
[b] N/A, not applicable.

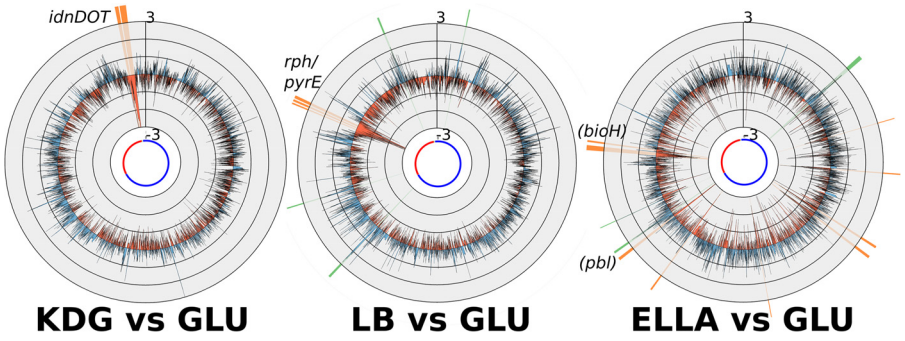

FIG 3 GLINT-normalized fitness profiles across the genome for the three conditions considered, projected onto the MG1655 genome. In each case, the selection score is based on the log ratio of transposon frequencies under the reference and selective conditions; thus, a positive score indicates that the presence of a genomic segment from the donor strain is detrimental, relative to the corresponding region in the host strain, under the selective condition. All scores are corrected for neighboring homology as described in Materials and Methods and are smoothed with a running median over a 1-kb window. Green and orange wedges correspond to windows flagged as significant by GLINT with positive and negative mean scores, respectively. The small chromosomal schematic in the center of each plot shows coloring corresponding to the conserved regions in Fig. S1 in the supplemental material. Labeled peaks were investigated in follow-up experiments, as described in the text; each is referred to by the locus ultimately identified as being responsible for the phenotypic differences or by the name of a nearby gene (if in parentheses, the locus is used only as a landmark). KDG, MOPS-KDG; GLU, MOPS-glucose; ELLA, MOPS-glucose-ellagic acid.

transfer of the *idnDOT* region via P1*vir* transduction from MG1655 to ATCC 8739 (marked with a kanamycin cassette between *yjgB* and *leuX*) yields a strain that grows rapidly in MOPS-KDG; however, when the same region, with *idnDOT* replaced with a kanamycin resistance cassette (Δ*idnDOT*), is transferred, the recipient strain ATCC 8739 remains unable to grow on KDG (Fig. 4B and C). Thus, GLINT immediately identifies the location of the single gene causing the differing growth phenotypes of MG1655 and ATCC 8739 on 5-keto-D-gluconate as a sole carbon source. We attribute the fact that ATCC 8739 cells with MG1655 *idnDOT* grow faster than MG1655 cells in MOPS-KDG medium to the fact that ATCC 8739 in general grows faster than MG1655 in minimal medium due to pyrimidine starvation in the latter (see below).

The exact peak positioning observed in this case, offset slightly from the actual location of the locus causing an actual fitness difference, is characteristic of patterns that we have observed in both ADAM and GLINT data. The dominant signal in the selection score is not typically centered on the gene of interest but rather appears one or a few kilobases away in either or both directions. This likely occurs because any transposon insertion inside the gene of interest would inactivate that gene in any case and thus yield no signal; in the absence of a signal from the precise location of the gene causing a fitness difference, asymmetry about that point can easily arise if there is any difference in the density of transposon insertions in the secondary library, or in the distance to the nearest insertion, on one side of that site versus the other. Thus, researchers applying global linkage analysis techniques must not assume that the gene at the center of the selection score peak is directly responsible for the phenotype observed; rather, they should inspect the identities of genes located within a few kilobases of the peak to determine the most likely contributors. Targeted follow-up experiments on the region immediately surrounding GLINT peaks, such as those described in the sections below, may become necessary if the identity of the sequence difference responsible for a selection score peak cannot be determined on the basis of available annotations.

TABLE 2 Peak locations identified by GLINT under each condition

Selection[a]	Peak location		Mean selection score[b]	Label[c]
	Start position	End position		
KDG vs GLU	4494480	4511380	−2.227	*idnDOT*
	4516980	4546830	−3.317	*idnDOT*
LB vs GLU	151980	157280	1.788	NA
	2874580	2884980	2.153	NA
	3259880	3265730	1.556	NA
	3760230	3772230	−2.55	*rph-pyrE*
	3779080	3788830	−2.358	*rph-pyrE*
	3794780	3818430	−4.755	*rph-pyrE*
	4349780	4357980	1.673	NA
ELLA vs GLU	1086830	1092830	−1.854	NA
	1211130	1220230	−2.007	NA
	2185430	2191080	−1.900	NA
	2985880	2993880	−2.528	*pbl*
	3029730	3035180	1.945	*pbl*
	3550080	3556630	−1.874	*bioH*
	3649480	3656580	−1.849	NA

[a] See Table 1, footnote *a*.
[b] Average of the rolling median statistic over the extent of the peak.
[c] NA, not applicable.

Comparison of growth in MOPS-glucose and LB medium. Perhaps the most obvious phenotypic difference between MG1655 and ATCC 8739 cells is not their difference in growth under exotic conditions but their widely disparate growth rates under standard conditions, particularly in minimal medium. ATCC 8739 cells grow approximately 10% faster in rich medium (LB) and 30% faster in glucose minimal medium (MOPS-GLU) than MG1655 cells (Fig. 5C). The relatively higher fitness (assessed using relative growth rates) of MG1655 in rich medium allowed us to investigate the genetic basis for the substantial difference in growth rates between the two strains in glucose minimal medium; as seen in Fig. 3 and 5A, GLINT profiling shows that the effects are dominated by a single site in the vicinity of the *waaQGPSBIJYZU* operon (giving rise to three closely spaced peak calls). In order to identify the specific gene responsible for the difference, we introduced a kanamycin resistance cassette into MG1655 adjacent to the *waa* region and then analyzed the growth of four independent

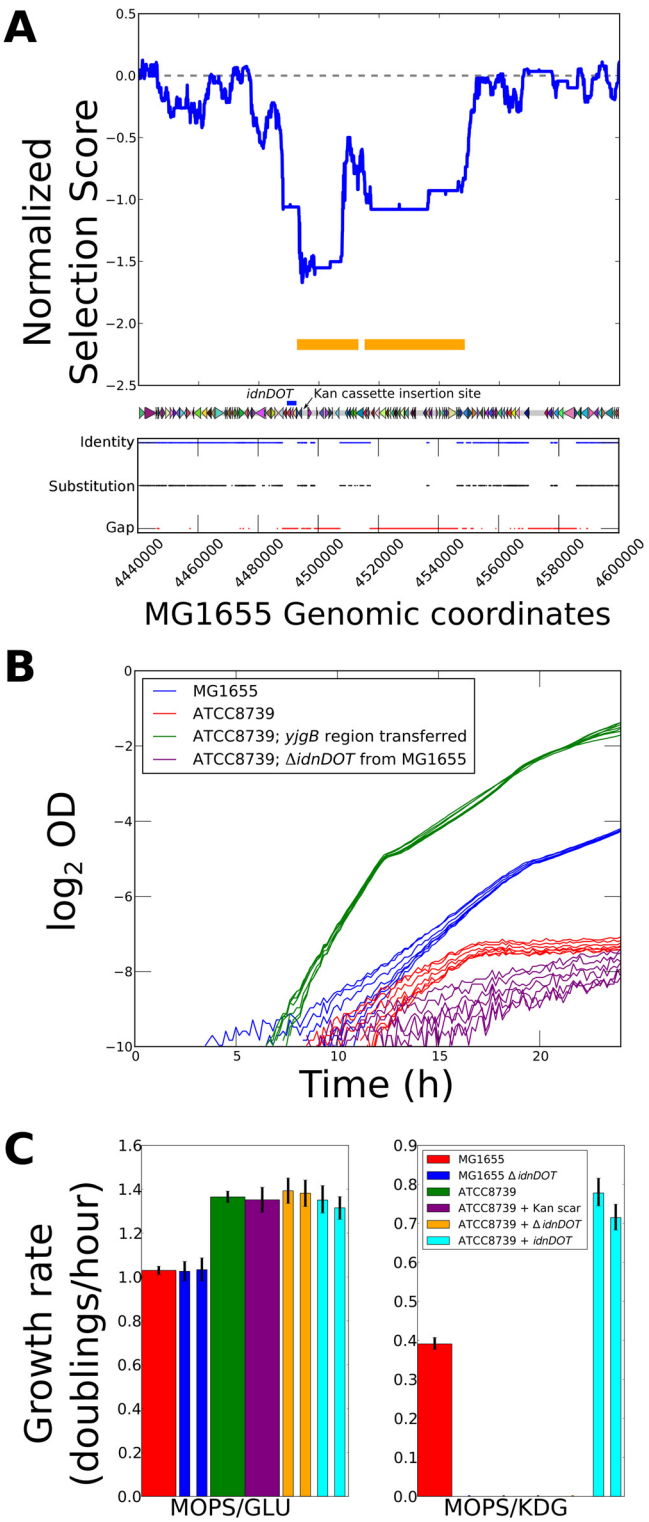

FIG 4 GLINT results for selections using 5-keto-D-gluconate (KDG) as a carbon source. (A) Normalized selection scores (averaged with a running median over a 5-kb window), genome schematic, and conservation status between MG1655 and ATCC 8739 for the region surrounding the dominant peak from the KDG selections. Negative selection scores indicate a lower fitness for the ATCC 8739 variant at a given position. (B) Growth curves in MOPS-KDG for the two baseline strains in this study, as well as for ATCC 8739 with a genomic region transferred from MG1655 by using either a Kan cassette downstream of *yjgB* (+ *idnDOT*) or a cassette knocking out *idnDOT* (Δ*idnDOT*). (C) Growth rates in MOPS-glucose or MOPS-KDG for the baseline strains or strains with modifications to this genomic region. Growth rates were calculated as described in Text S1 in the supplemental material; error bars indicate 95% confidence intervals obtained by resampling the posterior distribution of model parameters. Twin bars indicate the presence of two independently transduced strains of a given genotype; a single broad bar indicates that only a single strain of a given genotype was used. Growth rates of zero are assigned to strains that failed to reach an OD of 2.0×10^{-6} within 48 h of inoculation.

ATCC 8739 transductants containing this region (simply labeled transductants A to D in Fig. 5B). Of these, strain A showed growth rates nearly indistinguishable from those of ATCC 8739, whereas strains B and C showed growth rates equivalent to that of MG1655 in minimal medium and midway between those of ATCC 8739 and MG1655 in rich medium (strain D showed atypical growth curves with a long lag time preceding steady-state growth, and thus, its growth rate cannot be compared directly to those of the other strains in our study; this variant was omitted from further analysis). We sought to identify the specific gene responsible for the fitness difference by amplifying a set of small genomic windows from the vicinity of the resistance cassette insertion site from each of the transduced strains (A to D) and determining whether each window originated from MG1655 or from ATCC 8739 (either by sizing or by Sanger sequencing of the PCR products). The results, shown in Fig. 5B, reveal that the gene responsible for the different growth of our baseline strains must be between positions 3813150 and 3823233 (in MG1655 numbering), since this is the only region of MG1655-specific DNA shared by chimeras B and C but not by chimera A. This range excludes the *waa* operon but includes the *rph-pyrG* region. MG1655 is known to contain a frameshift in the *rph* gene that exerts polar effects on *pyrE*, leading to reduced growth due to pyrimidine deficiency (32, 33), which could easily yield the phenotypic difference observed. Consistent with this interpretation, supplementation of MOPS-GLU medium with uracil (20 μg/ml) equalizes the growth rates of ATCC 8739 and all three chimeric strains and shows a proportionally far greater effect on MG1655 and the chimeras (B and C) containing the *rph* frameshift than on ATCC 8739 (Fig. 5C and D).

Effects of ellagic acid on MG1655 and ATCC 8739. Ellagic acid is a naturally occurring tannin that has been shown to inhibit bacterial DNA gyrase *in vitro* (34) and exhibits modest antibacterial activity *in vivo* (34, 35). In both our initial screens and follow-up experiments, we found that exposure to sublethal concentrations of ellagic acid (ELLA) causes a lag of several hours in the growth of both MG1655 and ATCC 8739 cells but that the lag for any given concentration of ellagic acid is much longer in MG1655 (Fig. 6D). Our application of GLINT to identify the genetic basis for the differing susceptibilities of the strains to ellagic acid identified several regions with significant, and roughly equal, contributions (Fig. 3). Surprisingly, in light of the actual differences in growth in an ellagic acid-containing medium between MG1655 and ATCC 8739 (which strongly favor ATCC 8739), we identified loci with fitness effects (assessed by GLINT) in either direction (that is, those predicted to increase or decrease the fitness of ATCC 8739 upon transfer from MG1655). Detailed characterizations of two such cases are shown below. One should note that the condition referred to here as "ellagic acid" is in fact a compound stress condition, comprising both ellagic acid and a high-pH medium to avoid precipitation (see Materials and Methods for details); we

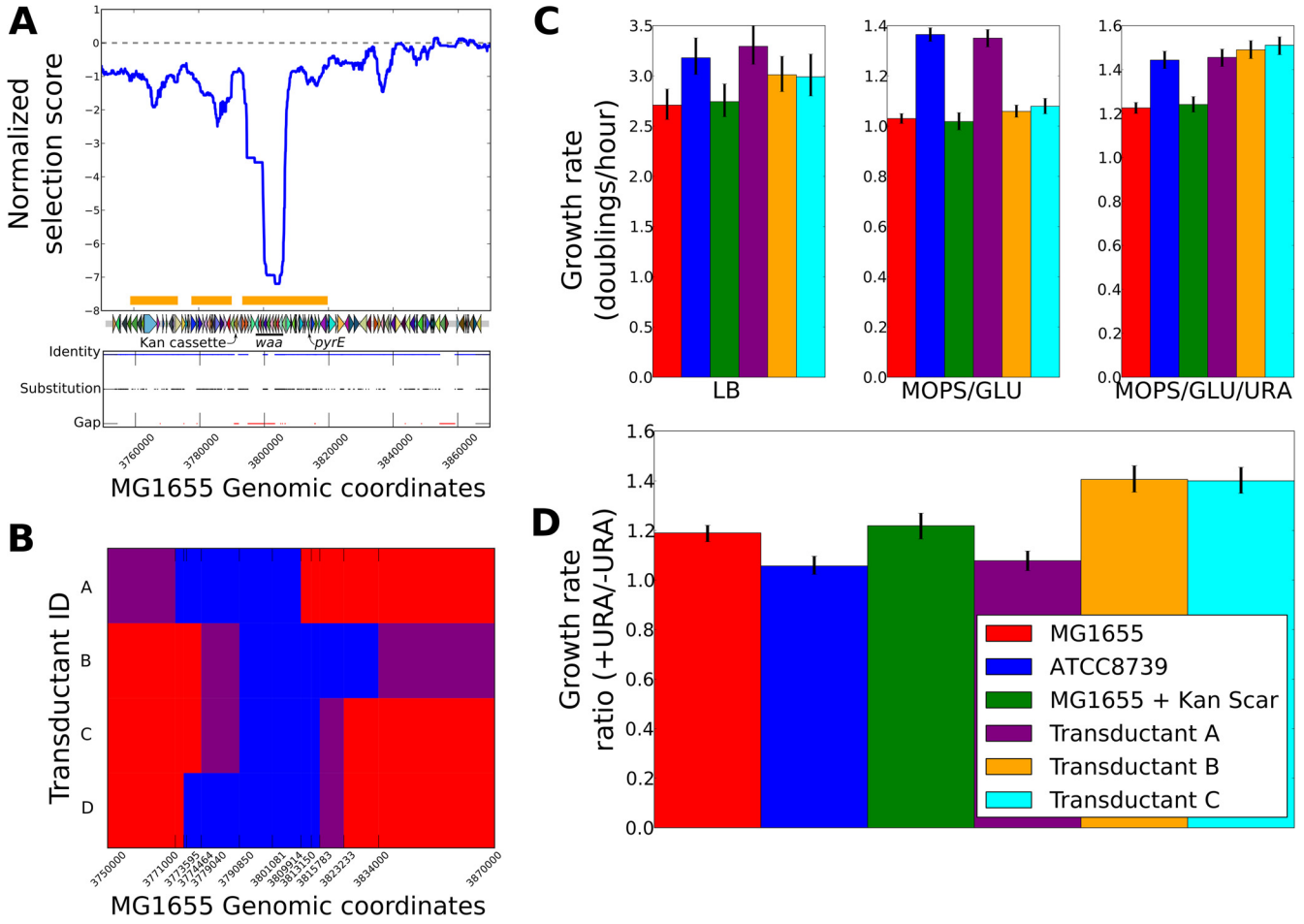

FIG 5 Effects of the MG1655 *rph* allele on the growth of ATCC 8739 cells. (A) Normalized selection scores (averaged over 5-kb windows) in the vicinity of the *pyrE* peak in the "LB vs GLU" selection (see Table 1), along with the locations of sequence mismatches between ATCC 8739 and MG1655. Orange bars under the selection scores show the extent of peak calls in this region. The region labeled *waa* contains *waaQGPSBIJYZU*. (B) Origins of genomic material for the region shown in panel A in each of four independently transduced ATCC 8739 derivatives containing the *pyrE* region from MG1655 (tagged with a kanamycin resistance cassette between *htrL* and *rfaD*). Ticks along the *x* axis show points at which identity was evaluated (by PCR or sequencing); regions of MG1655 origin are shown in blue, regions of ATCC 8739 origin in red, and regions where uncertainty exists (due to being between tested locations) in purple. The numbering corresponds to the MG1655 genome for consistency with panel A. (C) Growth rates of MG1655, ATCC 8739, and the ATCC 8739 derivatives containing the *pyrE* region from MG1655 (transductants A to C). Error bars indicate 95% confidence intervals based on resampling of the posterior model distribution. (D) Ratio of growth rates in MOPS-glucose plus uracil to those in MOPS-glucose for each of the strains for which results are shown in panel C. The color key applies to panels C and D.

refer to this condition as MOPS-GLU-ELLA for the sake of brevity.

The *bioH* region. The GLINT profiles in ellagic acid show a large negative peak in the vicinity of *bioH* (Fig. 6B), suggesting that transfer of this region from MG1655 to ATCC 8739 would increase fitness. Indeed, we constructed two test strains in the ATCC 8739 background where the *bioH* region had been transferred from MG1655 and found that both chimeric strains showed significantly shorter lag phases in MOPS-GLU-ELLA (Fig. 6) than ATCC 8739 or MG1655. While the chimeric strains consisting of ATCC 8739 with MG1655 *bioH* also show somewhat lower maximal growth rates (Fig. 6E), the difference in lag times is sufficient to give the chimeric strains a competitive advantage for at least 15 doublings (more than enough to saturate the culture in our selections, and almost any other conceivable environment). Unlike the KDG and LB cases discussed above, it is not clear from our findings what feature of this region conveys an advantage to the ATCC 8739 cells that receive it (we refer to this region as the *bioH* region simply as a landmark, not to imply some causal role of *bioH* itself). There are six small MG1655- and ATCC 8739-specific regions within a 50-kb window centered on this peak, any of which might provide the growth phenotype observed here. In a GLINT application where an investigator wishes to determine precisely what genetic difference between the two strains gives rise to the fitness effect observed, more-detailed analysis is straightforward; it can be done either through the analysis of the exact transfer boundaries of several transductants of this region (as in our investigation of the LB case, above, and the *pbl* region, below) or by constructing strains in the recipient background with each recipient-specific region deleted, and transferring strains from the donor to the recipient in which each donor-specific region is replaced by a selective marker. Phenotypic characterization of either set of strains would provide a more precise location for the genetic difference responsible.

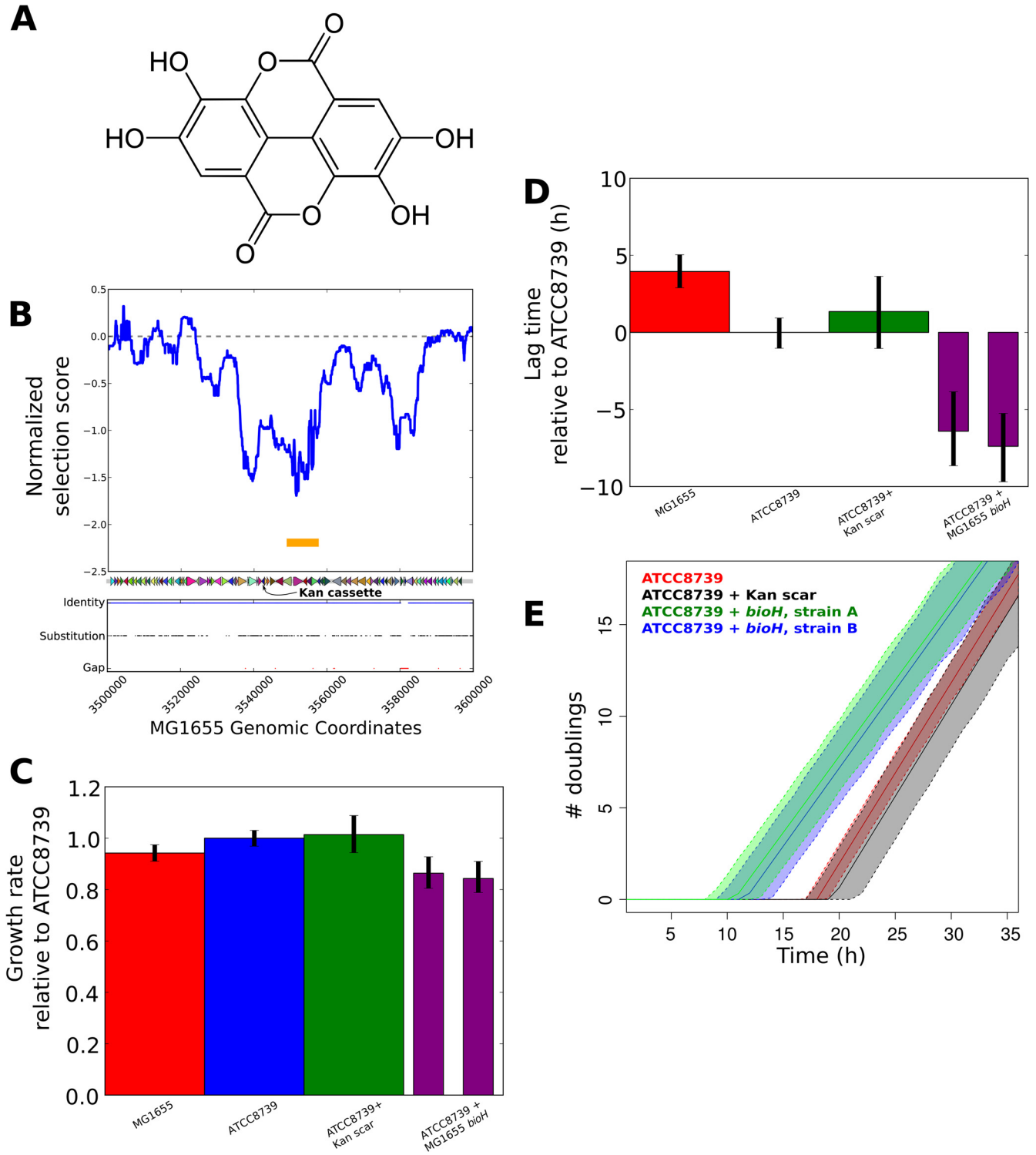

FIG 6 Growth of strains with the *bioH* region transferred in MOPS-glucose-ellagic acid medium. (A) Structure of ellagic acid (public domain image). (B) Normalized selection scores (averaged over 5-kb windows) in the vicinity of *bioH*. The location of the kanamycin cassette used to transfer the region is shown. GLINT peak calls are shown by orange bars. (C) Maximum growth rates of the two wild-type strains in MOPS-glucose-ellagic acid medium, compared with those of two independent transductants in which the *bioH* region of MG1655 was transferred to ATCC 8739. Growth rates are scaled relative to the average for ATCC 8739 replicates from the same day. (D) Lag times in MOPS-glucose-ellagic acid medium for the strains for which results are shown in panel C. All times are offset relative to the average lag for ATCC 8739 replicates from the same day. (E) Model-based growth curves for ATCC 8739 versus ATCC 8739-plus-MG1655 *bioH* transfer strains upon transition into a medium containing ellagic acid. Using the definitions provided in the text, strains are assumed to show no growth until their lag times and then to grow at their maximum specific growth rate for all subsequent times. The solid lines show the results from model fits; dashed lines show 95% confidence intervals (constructed for each strain using results from 1,000 draws of growth rates and lag times from the posterior distribution).

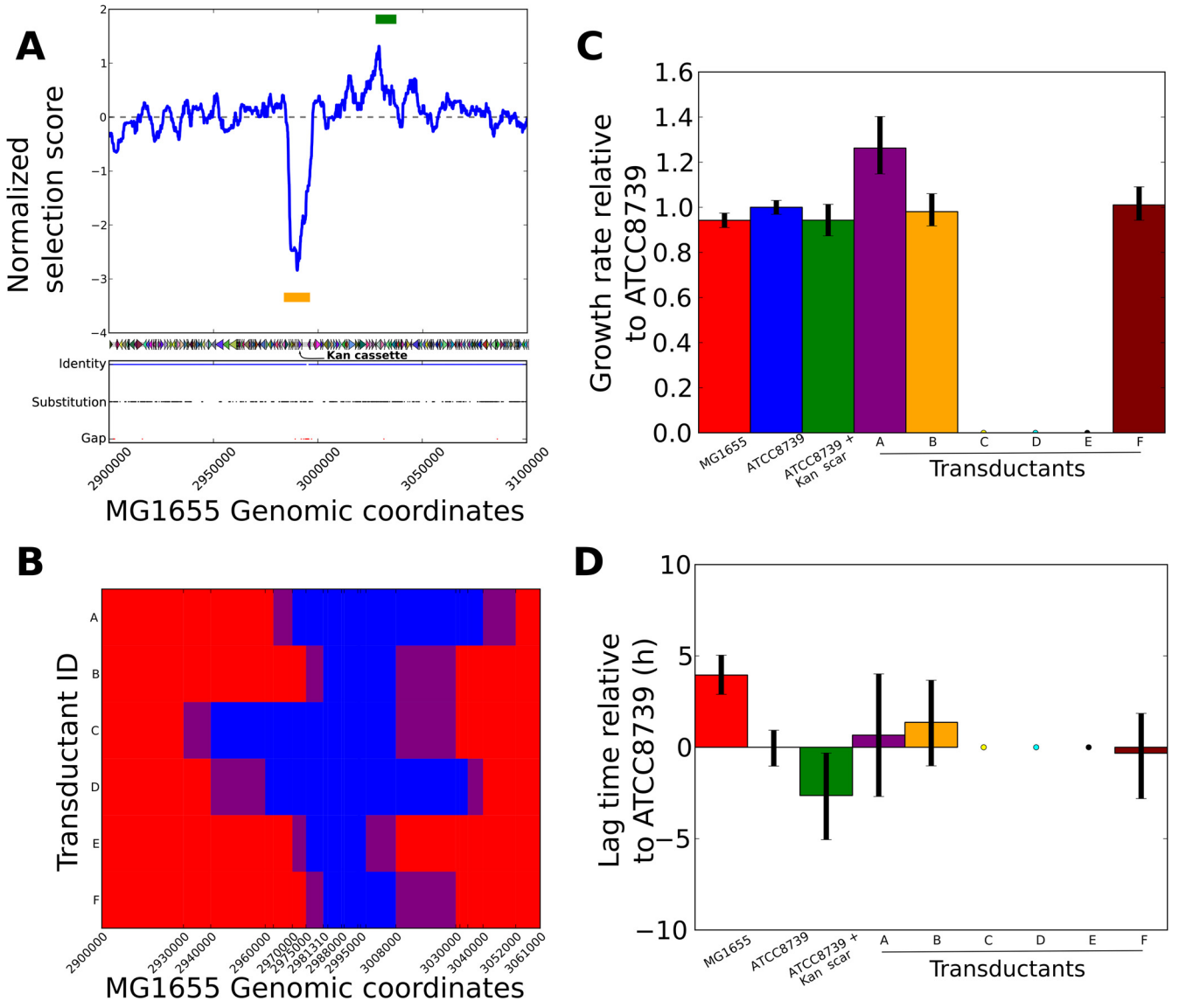

FIG 7 Growth of strains with the *pbl* region transferred in MOPS-glucose-ellagic acid medium. (A) Normalized selection scores in the vicinity of *pbl* (averaged over 5-kb regions). Orange and green bars mark negative and positive peak calls, respectively. (B) Strain of origin of genetic material in the region shown in panel A in each of six independently transduced ATCC 8739 derivatives (strains A to F) containing MG1655 DNA marked by a kanamycin resistance cassette upstream of *pbl*. Ticks along the *x* axis show points at which identity was evaluated; regions of MG1655 origin are shown in blue, regions of ATCC 8739 origin in red, and regions where uncertainty exists in purple. Note that the abscissas of these panels are not aligned. (C) Maximum growth rates of the two wild-type strains in MOPS-glucose-ellagic acid medium, compared with those of the transductants diagramed in panel B. Dots at point 0 are shown for strains that did not grow sufficiently for fitting during a 48-h experiment. Growth rates are scaled relative to the average for ATCC 8739 replicates from the same day. (D) Lag times in MOPS-glucose-ellagic acid medium for the strains for which results are shown in panel C. All times are offset relative to the average for ATCC 8739 replicates from the same day. Dots at point 0 are shown for strains that did not grow sufficiently for fitting during a 48-h experiment. Because the antibiotic resistance scar also had a measurable effect on lag times in the ATCC 8739 background, growth data for a strain containing this scar but no transfer of genomic material from MG1655 are also shown.

The *pbl* region. Unlike all other peaks identified in the course of this study, one region of the ellagic acid selection showed a pair of significant peaks in opposite directions directly adjacent to each other: a negative peak (one for which transfer from MG1655 to ATCC 8739 is expected to be beneficial) at positions 2985880 to 2993880 (8 kb) and a positive peak (one for which transfer from MG1655 to ATCC 8739 should be deleterious) from positions 3029730 to 3035180 (5.5 kb) (Fig. 7A). To characterize the effects of transferring this region of the genome to ATCC 8739, we placed a kanamycin resistance cassette upstream of the *pbl* pseudogene and generated six independently transduced strains in which this region was moved to the ATCC 8739 background (the boundaries of the transferred regions for these six strains are shown in Fig. 7B); the resulting chimeric strains are arbitrarily labeled with the letters A through F. Three of the resulting strains, strains C to E, show essentially no growth for 36 to 48 h upon transfer to a medium containing ellagic acid (substantially worse than the growth of the parental strain); one (strain A) shows a significantly

higher growth rate than the parental strain in this medium but no change in lag time; one (strain B) shows a slightly increased lag time relative to the strain with the scar alone; and the last (strain F) does not differ significantly from wild-type ATCC 8739 or the strain with the transduction scar alone in either growth rate or lag time. Thus, the presence both of an allele near *pbl* causing a decrease in fitness in ellagic acid medium upon transfer from MG1655 to ATCC 8739 and of a separate allele causing a substantial fitness increase are confirmed.

Analysis of the locations of transferred genomic material shown in Fig. 7B permits more-detailed identification of the genes responsible for this behavior. The only section of transferred DNA unique to strain A (which showed enhanced growth) was between positions 3034395 and 3052000 in the MG1655 genome. The strains rendered unable to grow in ellagic acid medium (strains C to E) share some portion of the region between positions 2970000 and 2981310, as does strain A, whereas the neutral strains (B and F) do not. Both of these regions are largely homologous, but not identical, between MG1655 and ATCC 8739. The most parsimonious explanation for these observations is that the MG1655 allele for some gene between positions 3034395 and 3052000 provides a substantial growth benefit in ellagic acid medium when transferred into the ATCC 8739 background, and the MG1655 allele for some gene between positions 2970000 and 2981310 is deleterious upon transfer. The beneficial allele must suppress the loss of growth caused by the deleterious allele, given the growth behavior of strain A. If desired, a more-detailed analysis could be performed by knockouts or targeted transfers of individual genes of interest in these regions, but we did not do so for the present application (the most obvious targets, based on relatively low sequence identity and annotated functionality, would be *ygfZ* for the beneficial MG1655 allele and *aas* for the deleterious MG1655 allele). Our analysis presented above assumes that during the transduction, only a single, contiguous block of the genome was transferred between MG1655 and ATCC 8739; we have observed a single case where multiple recombination events lead to the incorporation of two discontinuous, but nearby, sections of the donor genome (P. L. Freddolino, unpublished data), and thus, it is also, in principle, possible that some of our observations are explained by the presence of multiple crossovers that could be revealed only by further follow-up experiments.

The situation in the *pbl* region deserves particular attention regarding the interpretation of GLINT selection scores, given that it contains a strong negative peak and a weaker positive peak in close proximity (separated by ~35 kb). The results described in the preceding paragraph indicate that the effects of the beneficial MG1655 allele (negative selection score) suppress those of the deleterious MG1655 allele, consistent with the magnitude of the positive peak in this region being lower than expected for such a strong phenotypic effect (that is, lack of viability under the selective condition). While the positive peak here nevertheless passed the final size filter imposed by GLINT, the fact that it is relatively narrow means that it is less robust to user decisions regarding peak-calling thresholds than the other selection score peaks discussed above (see Text S2 in the supplemental material for further information). Users of GLINT in other contexts should be aware that in the (presumably rare) case of nearby loci with strongly opposed effects, one of the two may show partly suppressed selection scores due to the other, which may necessitate examination of the raw (rather than merged and filtered) significant probe calls if unexpected results are observed during follow-up experiments. These intermediate significance calls are also output by the GLINT software.

Effects of homology on selection scores. The principal difference between the application of global linkage analysis described here and previous applications such as ADAM (8) is the presence of substantial variations in the level of homology of the donor and recipient strains throughout the genome; thus, it is crucial to analyze the effects of homology on the resulting selection scores and to ensure that any biases are minimized during postprocessing.

Spearman correlations between the normalized selection score data and local homology scores (using a variety of averaging window sizes for the homology) are shown in Fig. S2A in the supplemental material, alongside similar results from a recent ADAM experiment performed using identical methods (data from reference 9). The striking pattern that emerges is that while there is no monotonic correlation between selection score and homology, a significant negative correlation exists in all cases between the local homology and the magnitude of the selection score—that is, low-homology regions show larger signals (no such correlation is observed in the ADAM results, where the transfer is between strains with no more than few dozen base pairs changed between them). Investigation of the distribution of selection scores at different levels of local homology shows that the distribution in all cases remains roughly symmetric and centered at zero even for low homologies, and thus, even for low-homology regions, the majority of transfers show no significant effect.

More detailed exploratory data analysis suggested that extended MG1655-specific portions of the genome showed particularly high variances in selection scores and, more importantly, high autocorrelations. Both observations can be easily understood as follows: in ADAM, where one strain of interest is a direct descendant of the other, a 1:1 mapping exists between essentially all positions on the donor and recipient genomes, and thus, each transposon insertion is representative of an equal-size chunk of the genome. In cases where a large insertion exists in the donor strain, however, any transposon insertions in the donor strain within that insertion will be indistinguishable in the recipient strain, and thus, when the transposon locations are mapped to the donor strain genome, the effects of all transposon insertions in a given donor-specific region should be highly correlated. In the nonnormalized GLINT selection score profiles, this phenomenon manifests itself as the presence of highly correlated, and often high-magnitude, scores in and around MG1655-specific regions. A schematized example of this phenomenon is shown in Fig. S3 in the supplemental material.

The above observations regarding the effects of donor-specific insertions suggested a highly effective set of corrections for the analysis of GLINT data: any donor-specific regions are mapped to a single location (to equalize the weight given to different transposon insertions), and then the magnitudes of scores as a function of the distance to the nearest large insertion are rescaled to standardize their distribution (to correct for the increased variances observed adjacent to MG1655-specific insertions). In practice, the latter step is done in GLINT by fitting a loess model to the probe-level distance-versus-selection score magnitude data and then dividing the raw selection score at each probe by the magnitude predicted at that homology by the loess model. As shown in Fig. S2B in the supplemental material, the correction applied is effec-

tive in eliminating the observed correlation between selection scores and local homology.

Despite the presence of some increased noise in low-homology regions, among the small number of examples that we studied in detail, we were able to identify fitness contributions from loci located in low-homology regions (the *idnDOT* operon in KDG medium described above), a point mutation adjacent to a low-homology region (the *rph* frameshift in glucose minimal medium), differences located in a high-homology region (the *bioH* region in ellagic acid-containing medium), and fitness differences due to nearby loci with opposed effects located in a high-homology region (the *pbl* region in ellagic acid-containing medium). Thus, with appropriate corrections, our method does not appear constrained to high- or low-homology regions and additionally is able to identify multiple strongly contributing loci when they exist.

DISCUSSION

Global linkage analysis, exemplified by the original ADAM method, provides a general-purpose method for identifying genetic differences between closely related organisms that give rise to selectable phenotypes. Here we provide a corresponding method, GLINT, for identifying similarly meaningful genetic differences between distantly related bacterial strains, using global linkage analysis followed by appropriate computational postprocessing. The method described here is highly modular, allowing it to be applied to a broad range of microbial systems. All that is required, in principle, are genome sequences for both strains of interest (of sufficient quality to align the genomes to each other), a method for generating a library of genomic DNA from one of the two strains, tagged with a selectable marker (the primary library), a method for transferring random portions of that library to the genome of the second strain (the secondary library), and a method for analyzing the abundances of different tagged library elements in different populations of the secondary library. In the present application, we used Tn*5* transposon insertions to generate the primary library, P1*vir* transduction to generate the secondary library, and two-color tiling microarrays to measure insert abundance. Depending on the species and strains of interest, a wide variety of other options could be chosen (for example, the use of a different transposon, such as mariner [36], for primary-library generation, the use of a method such as conjugation for secondary-library generation, or the use of high-throughput sequencing instead of microarrays to quantify marker abundance). The only strong constraints on the selection of specific methods for each of these steps are that the primary library must cover the donor strain's genome as broadly and uniformly as possible and the secondary library should contain replacements of homologous portions of the recipient strain's genome with the corresponding region of the donor strain. The software pipeline described here should be applicable, after minor preprocessing, to any global linkage analysis data set, since the primary considerations in its design are invariant for any choice of specific methods for implementing the workflow described above.

Under all three conditions that we studied, the transfer of genomic regions identified by GLINT provided strong phenotypes that were consistent with the direction predicted by the analysis. Nevertheless, in the GLU-versus-LB and ELLA-versus-LB cases, the particular regions on which we focused did not completely explain the differences in growth characteristics between strains, and other unexplored peaks do exist in the GLINT fitness score profiles (Fig. 3). Further contributors to the fitness differences between these strains could easily be identified by further follow-up experiments targeting other selection score peaks, or by a new GLINT experiment applying the same selections to a secondary library in which characterized points of difference between the strains had already been transferred (a procedure similar to the ADAM procedure followed to identify a pair of epistatically interacting mutations in reference 9). While it is not clear from the present data what is the weakest phenotypic difference (in terms of relative fitness) that can be identified using this method, equivalent footprinting methods in transposon-mutagenized libraries have enabled the detection of growth rate differences on the order of 10% (9) or antibiotic MIC changes of 1.5-fold (14). One should expect a similar level of sensitivity from GLINT. It should also be noted that GLINT, like ADAM, does not distinguish between phenotypic differences due to changing expression patterns and phenotypic differences due to changes in the functional molecules (protein or RNA) themselves; rather, the location of any genetic difference causing a phenotypic difference will be detected.

An inherent asymmetry in the design and execution of GLINT experiments should also be kept in mind: the secondary library, by construction, will in general contain cells with all but one small genomic region derived from the recipient. This means that for complex traits, cases where multiple alleles contribute independently or show epistatic interactions between the variants present in the recipient strain will be readily identified due to the large fitness effects of perturbing them, but epistatic interactions that are mechanistically present only between portions of the donor genome will not show their expected effects (except in the case of antagonistic epistasis within the donor genome, in which the variants present there may show their individual effects when placed in the recipient genome). Thus, for example, a trait that arises only due to synergistic interactions within the donor genome may not have its genetic basis properly mapped by GLINT. The possibility of new interactions between loci that arise only when a donor-specific variant is placed in the recipient genome must also be considered; in such cases, any phenotypic effects of the interaction will appear in the GLINT signal as centered on the donor locus that causes the interaction. As a practical matter, all of the above considerations suggest that the strain with the more "interesting" or extreme phenotype, if such a distinction can be made, should be the recipient strain in any GLINT pairing, since this will maximize the sensitivity of the method in detecting any recombination event that alters this phenotype.

The applicability of the method described here to other pairs of bacterial strains or species hinges primarily on the ability of the method chosen for the generation of the secondary library to transfer all portions of the donor genome into the recipient background: any insertions whose effects should be evaluated cannot be larger than the size of the region transferred (tens of kilobases, for P1 transduction [18]), and the strains under consideration must possess enough homology to allow homologous recombination at sites throughout the genome. At present, insufficient data are available to provide a hard cutoff for how distantly related strains may be to satisfy the latter condition. It is evident from our results that GLINT performs fairly efficiently even in the vicinity of strain-specific insertions and low-homology regions: the *idnDOT* insertion identified as crucial to growth on KDG, for example, is contained in a 16-kb window (in the MG1655 ge-

nome) in which only ~50% of bases match between MG1655 and ATCC 8739; in addition, *idnDOT* itself is part of a 5-kb MG1655-specific insertion and is adjacent to 8-kb and 29-kb MG1655-specific insertions. Nevertheless, as seen in Fig. 4A, GLINT provides a strong indicator of the presence of a key difference between the strains in this particular low-homology region. Extrapolating from this experience, it appears likely that P1 transduction is sufficiently efficient even for low-homology regions, so that even the presence of large strain-specific regions does not necessarily pose a barrier. The only regions for which effects on fitness differences are likely to be missed are those in insertions that are simply too large to be transferred at all by the secondary-library generation method in use (that is, around 90 kb for P1 transduction [18]). This constraint should be considered in the context of extremely distantly related strains that might be investigated using GLINT. For example, Welch et al. (37) showed that in pairwise comparisons with MG1655, a uropathogenic *E. coli* strain (CFT073) and an enterohemorrhagic *E. coli* strain (EDL933) show as few as 52.2% of predicted proteins in common, and only 39.2% are common to all three strains. However, in spite of this extremely low overall homology, only 5 to 10 islands specific to each strain are large enough to be nontransferable using P1 transduction (see Fig. 3 of reference 37). While we have not attempted to identify fitness differences arising between such distantly related strains, in principle we expect that any important alleles outside of these very large strain-specific regions should be detectable using GLINT, and the use of a less size limited secondary-library generation method might remove this barrier.

GLINT should thus be useful in comparisons even between very distantly related strains, provided that a few limitations are kept in mind. Ideally, investigators attempting to apply the method to very distantly related strains, or to a cross-genus comparison, should confirm (preferably by footprinting of transposon insertion locations) that the secondary library contains sufficiently broad coverage of the recipient strain that all potential regions of interest are close to one or more marker insertions (with "close" defined relative to the sizes of the genomic regions transferred during secondary-library generation). This is a fairly simple quality control step for the secondary library that should make it clear which genomic regions will not be covered by GLINT results. Even if such regions exist, GLINT will provide usable results for the remainder of the genomes under comparison. In such a case, the absence of a sufficiently strong difference identified in GLINT-comparable regions to account for an observed phenotypic difference could itself be taken as evidence that more attention must be focused on the very large insertions, ruling out contributions from the remainder of the genome.

One other possible complication that arises during the analysis of GLINT data is that transposon insertions are frequently not phenotypically neutral, and direct effects of insertions could in principle either mask an important phenotypic difference between strains or give rise to a false-positive result if an insertion were itself beneficial under the selective condition of interest. The former case (a transposon insertion masking the effects of an important allele) is unlikely to pose a problem in a sufficiently dense secondary library; given that P1 transduction spans linkage distances of tens of kilobases, there will be several distinct labeled strains in the secondary library near any particular locus of interest, and even if one or more of them contain a transposon insertion that masks the phenotype of interest, the effects of that locus will still be detectable from the other nearby strains. The possibility of a transposon insertion directly causing a change in fitness that is strong enough to appear in the GLINT fitness profile, however, does remain, and could in principle result in a false-positive result (although the signal due to such an insertion would still have to survive averaging with many other adjacent insertions that would show no such effect). We have never observed such a false-positive result in applications of either ADAM or GLINT, but the possibility of such an insertion underscores the need for validation of any loci identified by GLINT using follow-up experiments on independently generated chimeric strains.

A method conceptually similar to GLINT, extreme quantitative trait locus mapping (X-QTL) (38), has also recently been developed in *Saccharomyces cerevisiae*. X-QTL makes use of yeast genetics to generate a large pool of haploid progeny from a cross between two strains of interest and then applies high-throughput sequencing or a custom microarray to measure the allelic frequencies in this pool under selective versus unselective conditions. X-QTL and GLINT address similar problems in yeast and bacteria, respectively, using methods for generating libraries of hybrid strains appropriate to the organism being used; a few practical differences in the application and interpretation of the methods also exist. To our knowledge, there is no X-QTL analog of the computational postprocessing used in GLINT to correct for the effects of low-homology regions, although we have shown that such a correction is important for the comparison of naturally occurring strains of *E. coli*. We are not in a position to judge whether similar corrections would be useful in yeast populations. The methods will also differ substantially in their behavior surrounding complex polygenic traits. Consider a trait involving the interaction of n loci; in X-QTL, the hybrid library will contain strains with a roughly binomial distribution of the number of these loci obtained from each parental strain, and thus, library strains with 1 or $n - 1$ of those loci from one parent will be relatively rare. In GLINT, on the other hand, the library will contain almost entirely strains with either n or $n - 1$ alleles at these loci from the recipient strain, and zero or 1 from the donor strain. Thus, in principle, GLINT should provide a more-focused picture of the most strongly contributing alleles in the context of the recipient strain background, whereas X-QTL will provide information on the effects of each allele averaged over a distribution of genotypes at other contributing loci. Neither of these behaviors seems inherently superior (and because these behaviors arise due to the necessary genetic manipulations of the host rather than due to some design choice, they cannot be easily altered in any case), but they should be kept in mind when one is interpreting results obtained using either of these methods.

In summary, we have extended a previously developed global linkage analysis method to enable rapid identification of the genetic differences responsible for differing phenotypes of naturally occurring bacterial strains. Using GLINT, we are able to identify the genetic bases of both simple and complex traits, present in high- or low-homology regions of the genomes of distantly related strains. The applicability of this family of methods is limited only by the presence of suitable methods for generating the requisite tagged library in one strain of interest and for performing homologous transfer of genomic segments from the tagged library to the second strain. These methods will be useful in any case where researchers seek to identify the genetic basis of a selectable trait in a specific microbial population. Two particularly pertinent exam-

ples are the identification of specific genetic variations giving rise to different levels of antibiotic resistance in natural or clinical populations (39) and the optimization of strains for biosynthetic applications by improving their usage of particular carbon sources or their tolerance of toxic intermediates or by-products (23, 40, 41).

ACKNOWLEDGMENTS

This research was supported by grants from NIAID (5R01AI077562) and the NIH Director's Pioneer Award (8DP1ES022578).

We thank Anupama Khare for technical assistance in the construction of the primary transposon library.

REFERENCES

1. **Vignaroli C, Luna GM, Rinaldi C, Cesare AD, Danovaro R, Biavasco F.** 2012. New sequence types and multidrug resistance among pathogenic *Escherichia coli* isolates from coastal marine sediments. Appl. Environ. Microbiol. **78:**3916–3922. http://dx.doi.org/10.1128/AEM.07820-11.
2. **Knoshaug EP, Zhang M.** 2009. Butanol tolerance in a selection of microorganisms. Appl. Biochem. Biotechnol. **153:**13–20. http://dx.doi.org/10.1007/s12010-008-8460-4.
3. **Applebee MK, Herrgard MJ, Palsson BO.** 2008. Impact of individual mutations on increased fitness in adaptively evolved strains of *Escherichia coli*. J. Bacteriol. **190:**5087–5094. http://dx.doi.org/10.1128/JB.01976-07.
4. **Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF.** 2009. Genome evolution and adaptation in a long-term experiment with *Escherichia coli* mu. Nature **461:**1243–1247. http://dx.doi.org/10.1038/nature08480.
5. **Tenaillon O, Rodriguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, Gaut BS.** 2012. The molecular diversity of adaptive convergence. Science **335:**457–461. http://dx.doi.org/10.1126/science.1212986.
6. **Mulcahy LR, Burns JL, Lory S, Lewis K.** 2010. Emergence of *Pseudomonas aeruginosa* strains producing high levels of persister cells in patients with cystic fibrosis. J. Bacteriol. **192:**6191–6199. http://dx.doi.org/10.1128/JB.01651-09.
7. **Shea PR, Beres SB, Flores AR, Ewbank AL, Gonzalez-Lugo JH, Martagon-Rosado AJ, Martinez-Gutierrez JC, Rehman HA, Serrano-Gonzalez M, Fittipaldi N, Ayers SD, Webb P, Willey BM, Low DE, Musser JM.** 2011. Distinct signatures of diversifying selection revealed by genome analysis of respiratory tract and invasive bacterial populations. Proc. Natl. Acad. Sci. U. S. A. **108:**5039–5044. http://dx.doi.org/10.1073/pnas.1016282108.
8. **Goodarzi H, Hottes AK, Tavazoie S.** 2009. Global discovery of adaptive mutations. Nat. Methods **6:**581–583. http://dx.doi.org/10.1038/nmeth.1352.
9. **Freddolino PL, Goodarzi H, Tavazoie S.** 2012. Fitness landscape transformation through a single amino acid change in the Rho terminator. PLoS Genet. **8:**e1002744. http://dx.doi.org/10.1371/journal.pgen.1002744.
10. **Gunsalus IC, Hand DB.** 1941. The use of bacteria in the chemical determination of total vitamin C. J. Biol. Chem. **141:**853–858.
11. **Neidhardt FC, Bloch PL, Smith DF.** 1974. Culture medium for enterobacteria. J. Bacteriol. **119:**736–747.
12. **Darling AE, Mau B, Perna NT.** 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One **5:**e11147. http://dx.doi.org/10.1371/journal.pone.0011147.
13. **Girgis HS, Liu Y, Ryu WS, Tavazoie S.** 2007. A comprehensive genetic characterization of bacterial motility. PLoS Genet. **3:**1644–1660. http://dx.doi.org/10.1371/journal.pgen.0030154.
14. **Girgis HS, Hottes AK, Tavazoie S.** 2009. Genetic architecture of intrinsic antibiotic susceptibility. PLoS One **4:**e5629. http://dx.doi.org/10.1371/journal.pone.0005629.
15. **Silhavy TJ, Berman ML, Enquist LW. (ed).** 1984. Experiments with gene fusions. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
16. **Cherepanov PP, Wackernagel W.** 1995. Gene disruption in *Escherichia coli*: Tc^R and Km^R cassettes with the option of Flp-catalyzed excision of the antibiotic-resistance determinant. Gene **158:**9–14. http://dx.doi.org/10.1016/0378-1119(95)00193-A.
17. **Datsenko KA, Wanner BL.** 2000. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. Proc. Natl. Acad. Sci. U. S. A. **97:**6640–6645. http://dx.doi.org/10.1073/pnas.120163297.
18. **Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K (ed).** 2001. Current protocols in molecular biology. John Wiley & Sons, Inc, New York, NY.
19. **Weng L, Dai H, Zhan Y, He Y, Stepaniants SB, Bassett DE.** 2006. Rosetta error model for gene expression analysis. Bioinformatics **22:**1111–1121. http://dx.doi.org/10.1093/bioinformatics/btl045.
20. **Nichols RJ, Sen S, Choo YJ, Beltrao P, Zietek M, Chaba R, Lee S, Kazmierczak KM, Lee KJ, Wong A, Shales M, Lovett S, Winkler ME, Krogan NJ, Typas A, Gross CA.** 2011. Phenotypic landscape of a bacterial cell. Cell **144:**143–156. http://dx.doi.org/10.1016/j.cell.2010.11.052.
21. **Cleveland WS, Grosse E, Shyu WM.** 1992. Local regression models, p 309–376. *In* Chambers JM, Hastie TJ (ed), Statistical models in S. Wadsworth & Brooks/Cole, Pacific Grove, CA.
22. **Kuan PF, Chun H, Keleş S.** 2008. CMARRT: a tool for the analysis of ChIP-chip data from tiling arrays by incorporating the correlation structure. Pac. Symp. Biocomput. **2008:**515–526. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2862456/.
23. **Schwalbach MS, Keating DH, Tremaine M, Marner WD, Zhang Y, Bothfeld W, Higbee A, Grass JA, Cotten C, Reed JL, da Costa Sousa L, Jin M, Balan V, Ellinger J, Dale B, Kiley PJ, Landick R.** 2012. Complex physiology and compound stress responses during fermentation of alkali-pretreated corn stover hydrolysate by an *Escherichia coli* ethanologen. Appl. Environ. Microbiol. **78:**3442–3457. http://dx.doi.org/10.1128/AEM.07329-11.
24. **Benjamini Y, Hochberg Y.** 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B **57:**289–300.
25. **Buck MJ, Nobel AB, Lieb JD.** 2005. ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. Genome Biol. **6:**R97. http://dx.doi.org/10.1186/gb-2005-6-11-r97.
26. **Madar D, Dekel E, Bren A, Alon U.** 2011. Negative auto-regulation increases the input dynamic-range of the arabinose system of *Escherichia coli*. BMC Syst. Biol. **5:**111. http://dx.doi.org/10.1186/1752-0509-5-111.
27. **van Opijnen T, Bodi KL, Camilli A.** 2009. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat. Methods **6:**767–772. http://dx.doi.org/10.1038/nmeth.1377.
28. **Sims GE, Kim SH.** 2011. Whole-genome phylogeny of *Escherichia coli/Shigella* group by feature frequency profiles (FFPs). Proc. Natl. Acad. Sci. U. S. A. **108:**8329–8334. http://dx.doi.org/10.1073/pnas.1105168108.
29. **Bausch C, Peekhaus N, Utz C, Blais T, Murray E, Lowary T, Conway T.** 1998. Sequence analysis of the GntII (subsidiary) system for gluconate metabolism reveals a novel pathway for L-idonic acid catabolism in *Escherichia coli*. J. Bacteriol. **180:**3704–3710.
30. **Bausch C, Ramsey M, Conway T.** 2004. Transcriptional organization and regulation of the L-idonic acid pathway (GntII system) in *Escherichia coli*. J. Bacteriol. **186:**1388–1397. http://dx.doi.org/10.1128/JB.186.5.1388-1397.2004.
31. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL.** 2009. BLAST+: architecture and applications. BMC Bioinformatics **10:**421. http://dx.doi.org/10.1186/1471-2105-10-421.
32. **Jensen KF.** 1993. The *Escherichia coli* K-12 "wild types" W3110 and MG1655 have an *rph* frameshift mutation that leads to pyrimidine starvation due to low *pyrE* expression levels. J. Bacteriol. **175:**3401–3407.
33. **Soupene E, van Heeswijk WC, Plumbridge J, Stewart V, Bertenthal D, Lee H, Prasad G, Paliy O, Charernnoppakul P, Kustu S.** 2003. Physiological studies of *Escherichia coli* strain MG1655: growth defects and apparent cross-regulation of gene expression. J. Bacteriol. **185:**5611–5626. http://dx.doi.org/10.1128/JB.185.18.5611-5626.2003.
34. **Ohemeng K, Schwender C, Fu K, Barrett J.** 1993. DNA gyrase inhibitory and antibacterial activity of some flavones. Bioorg. Med. Chem. Lett. **3:**225–230. http://dx.doi.org/10.1016/S0960-894X(01)80881-7.
35. **Akiyama H, Fujii K, Yamasaki O, Oono T, Iwatsuki K.** 2001. Antibacterial action of several tannins against *Staphylococcus aureus*. J. Antimicrob. Chemother. **48:**487–491. http://dx.doi.org/10.1093/jac/48.4.487.
36. **Lampe DJ, Churchill ME, Robertson HM.** 1996. A purified mariner transposase is sufficient to mediate transposition in vitro. EMBO J. **15:**5470–5479.
37. **Welch RA, Burland V, Plunkett G, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HLT, Donnenberg MS, Blattner FR.** 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. **99:**17020–17024. http://dx.doi.org/10.1073/pnas.252529799.

38. **Ehrenreich IM, Torabi N, Jia Y, Kent J, Martis S, Shapiro JA, Gresham D, Caudy AA, Kruglyak L.** 2010. Dissection of genetically complex traits with extremely large pools of yeast segregants. Nature **464:**1039–1042. http://dx.doi.org/10.1038/nature08923.
39. **Amini S, Tavazoie S.** 2011. Antibiotics and the post-genome revolution. Curr. Opin. Microbiol. **14:**513–518. http://dx.doi.org/10.1016/j.mib.2011.07.017.
40. **Yomano LP, York SW, Ingram LO.** 1998. Isolation and characterization of ethanol-tolerant mutants of *Escherichia coli* KO11 for fuel ethanol production. J. Ind. Microbiol. Biotechnol. **20:**132–138. http://dx.doi.org/10.1038/sj.jim.2900496.
41. **Yomano LP, York SW, Zhou S, Shanmugam KT, Ingram LO.** 2008. Re-engineering *Escherichia coli* for ethanol production. Biotechnol. Lett. **30:**2097–2103. http://dx.doi.org/10.1007/s10529-008-9821-3.