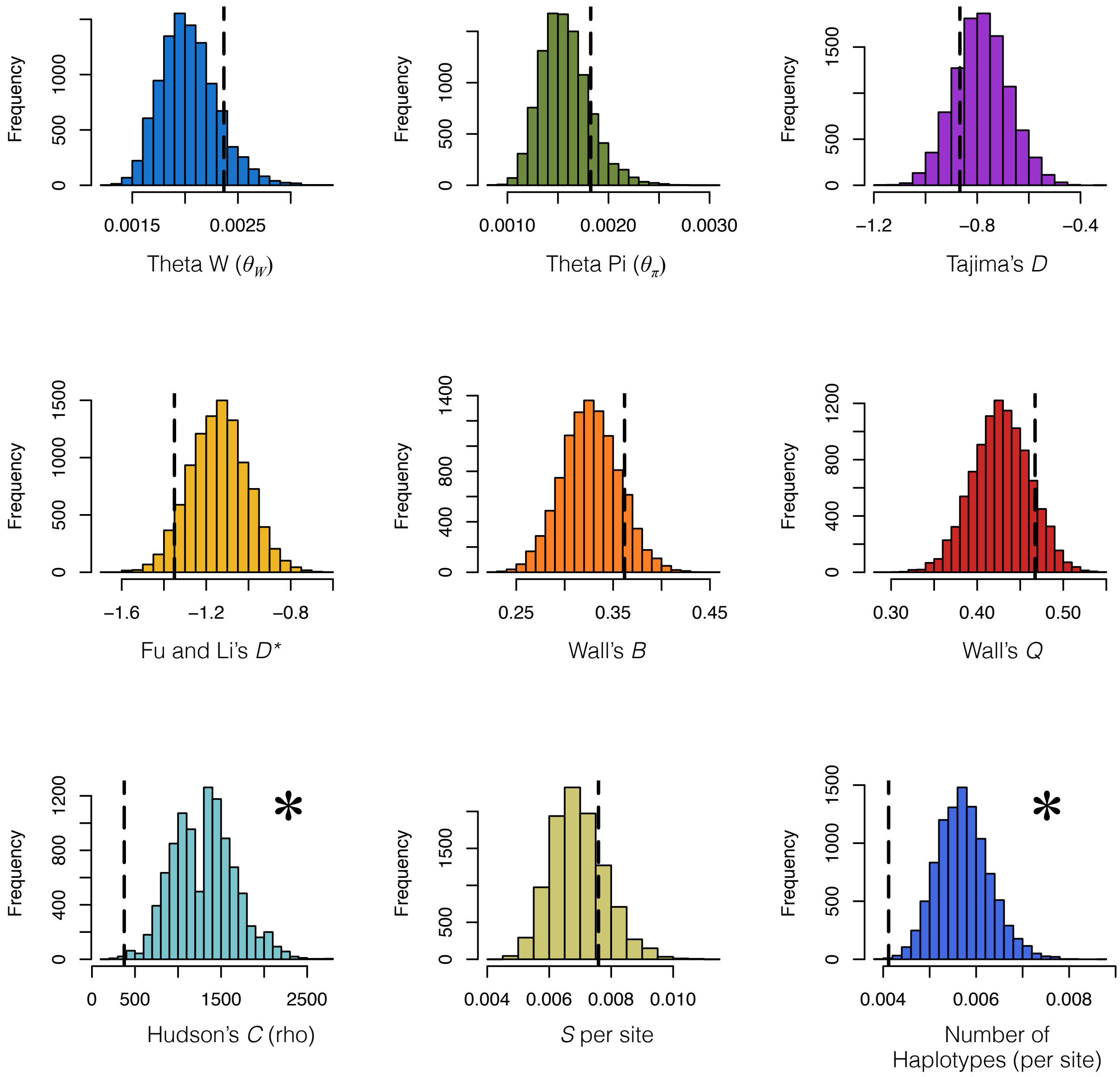

Figure S1, related to Figure 3. Difference between candidate genes versus random gene sets for different evolutionary metrics. Dotted lines indicate the mean value for each molecular evolution statistic for candidate genes as a group. The observed value for the group is compared to a null distribution based on 10,000 random gene sets of the same size. Asterisks indicate statistically significant results.

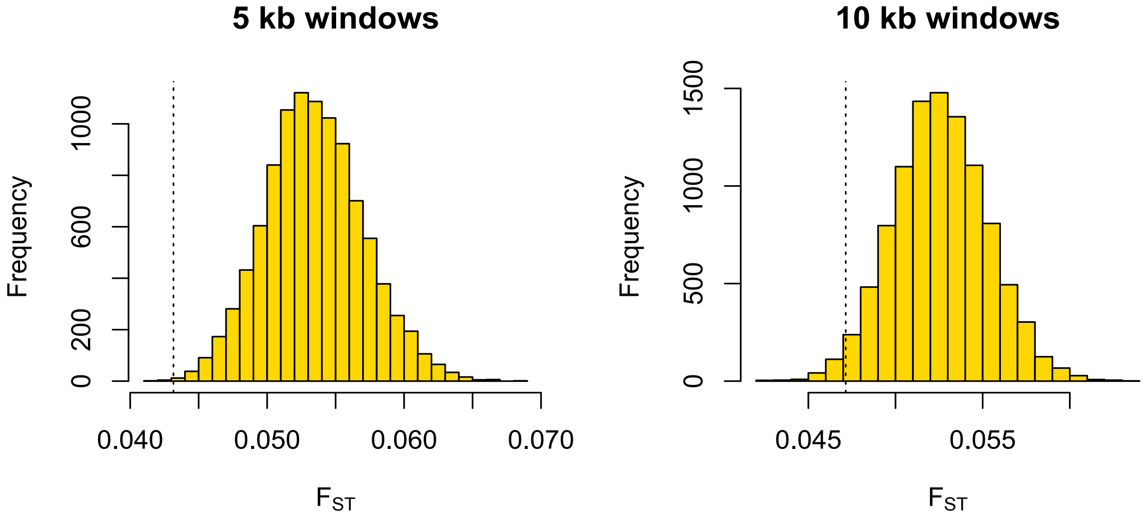

Figure S2, related to Fig. 3. Candidate gene regions show lower population structure (F_{ST} values) compared to randomly chosen regions of the genome. Dotted lines show the observed mean value of F_{ST} for candidate gene regions, whereas the distributions show the mean for 10,000 datasets of the same size based on randomly chosen regions in the genome. F_{ST} was calculated for all bi-allelic SNPs that were present in at least one of the two populations (Houston, TX or Mountain Lake, VA).

Metric	Observed Value for Candidate Loci	Lower 5th Percentile for Random Datasets	Upper 95th Percentile for Random Datasets
p_N	0.0008	0.0004	0.0010
p_S	0.0028	0.0020	0.0035
Within-species omega (p_N/p_S)	**0.200**	**0.079**	**0.196**
p_N/(p_N+p_S)	**0.217**	**0.135**	**0.192**
Number of times p_N>p_S[a]	**39.8**	**23.1**	**36.3**
d_N	0.054	0.045	0.071
d_S	**0.754**	**0.570**	**0.741**
***D. discoideum-D. citrinum* omega (d_N/d_S)**	**0.068**	**0.071**	**0.101**
P_N/D_N	0.117	0.042	0.182
P_S/D_S	**0.095**	**0.042**	**0.085**
MK-ratio[b]	1.034	0.931	1.504

[a] Out of 210 possible pairwise comparisons of 21 sequences (n[n-1]/2 = 21*20/2 = 210). This metric is similar to p_N/p_S but does not result in division by zero when p_S is zero.

[b] MK-ratio = (D_N/D_S)/(P_N/P_S). A value of '1' was added to all values in the 2x2 matrix to prevent division by zero.

Table S1. Levels of nonsynonymous (p_N) and synonymous (p_S) polymorphisms within *D. discoideum*, as well as nonsynonymous (d_N) and synonymous (d_S) substitutions between *D. discoideum* and *D. citrinum.* Mean value for candidate genes is shown in comparison to the 5th and 95th percentiles of the null distribution based on 10,000 datasets of the same size where the genes were chosen randomly.

A

Timing of Expression	Number of Candidate Genes (Proportion)	Number of Non-Candidate Genes (Proportion)
Unicellular Only	1 (0.01)	171 (0.01)
Multicellular Only	20 (0.23)	3487 (0.28)
Unicellular and Multicellular	43 (0.49)	5413 (0.43)
Not Expressed	23 (0.26)	3484 (0.28)

B

	Observed value, Candidate Genes	Random datasets, 5th Percentile	Random datasets, 95th Percentile
Mean vegetative (unicellular) expression level	1.19	1.11	1.48
Mean developmental (multicellular) expression level	1.63	1.55	1.89

Table S2. Sequence polymorphism as a function of timing of expression during the life cycle or expression level. (A) The number of candidate versus non-candidate genes categorized as 'vegetative' (unicellular expression only), 'developmental' (multicellular expression only), 'both' (unicellular and multicellular expression), or 'not_expressed' according to Parikh et al. [S1]. There was no significant overrepresentation of candidate genes in any of these categories (Fisher's Exact test, *P*=0.65). (B) Mean expression levels of candidate genes compared to randomly chosen datasets of the same size at two stages of the life cycle, based on data in Parikh et al. [S1].

	Size of Sequence Window	Lower 5th Percentile for Random Datasets	Upper 95th Percentile for Random Datasets	Observed Value for Candidate Loci
S	**10 kb**	**62.6**	**72.0**	**72.9**
Singletons	**10 kb**	**34.0**	**39.9**	**40.3**
Number of Mutations	**10 kb**	**62.8**	**72.3**	**73.1**
Number of Haplotypes	10 kb	18.4	19.1	18.7
Haplotype Diversity	**10 kb**	**0.9826**	**1.0006**	**1.0011**
Wall's B	**10 kb**	**0.293**	**0.326**	**0.332**
Wall's Q	**10 kb**	**0.439**	**0.474**	**0.475**
Theta W	**10 kb**	**0.0018**	**0.0021**	**0.0021**
Theta Pi	*10 kb*	*0.0014*	*0.0016*	*0.0016*
Tajima's D	10 kb	-1.07	-0.92	-1.04
Fu and Li's D*	10 kb	-1.59	-1.38	-1.53
Fu and Li's F*	10 kb	-1.64	-1.41	-1.58
Hudson's Ĉ (rho)	*10 kb*	*14.0*	*19.8*	*14.3*
S	*20 kb*	*149.5*	*167.2*	*167.2*
Singletons	*20 kb*	*84.6*	*95.6*	*94.4*
Number of Mutations	*20 kb*	*149.9*	*167.7*	*167.7*
Number of Haplotypes	20 kb	19.6	19.9	19.9
Haplotype Diversity	*20 kb*	*0.992*	*1.000*	*1.000*
Wall's B	20 kb	0.301	0.325	0.317
Wall's Q	20 kb	0.451	0.477	0.468
Theta W	**20 kb**	**0.0022**	**0.0025**	**0.0026**
Theta Pi	**20 kb**	**0.0017**	**0.0020**	**0.0020**
Tajima's D	20 kb	-1.10	-0.98	-1.02
Fu and Li's D*	20 kb	-1.74	-1.59	-1.64
Fu and Li's F*	20 kb	-1.71	-1.55	-1.59
Hudson's Ĉ (rho)	20 kb	19.28	23.29	21.60

Table S3. Results for sequence windows of 10 kb (5 kb to each side) and 20 kb (10 kb to each side). Analyses compare the observed mean values for candidate genes as a group to the distribution of mean values in 10,000 random datasets of the same size as the observed dataset. Boldface indicates metrics where the mean of candidate genes was below the 5th or above the 95th percentile of the null distribution. Italics indicate where the observed value for candidate loci was below the 10th or above the 90th percentile of the null distribution. In this analysis, sites with missing data for one or more strains were retained in the analysis.

	Size of Sequence Window	Lower 5th Percentile for Random Datasets	Upper 95th Percentile for Random Datasets	Observed Value for Candidate Loci
S	**10 kb**	**30.2**	**36.7**	**38.4**
Singletons	**10 kb**	**16.5**	**21.4**	**22.0**
Number of Mutations	**10 kb**	**30.3**	**36.8**	**38.5**
Number of Haplotypes	10 kb	13.3	14.4	14.1
Haplotype Diversity	10 kb	0.8819	0.9200	0.9049
Wall’s B	**10 kb**	**0.289**	**0.335**	**0.357**
Wall’s Q	**10 kb**	**0.402**	**0.455**	**0.471**
Theta W	**10 kb**	**0.00086**	**0.00104**	**0.00109**
Theta Pi	**10 kb**	**0.00064**	**0.00078**	**0.00080**
Tajima’s D	10 kb	-1.04	-0.84	-1.00
Fu and Li’s D*	10 kb	-1.53	-1.26	-1.47
Fu and Li’s F*	10 kb	-1.61	-1.33	-1.55
Hudson’s Ĉ (rho)	**10 kb**	**20.27**	**282.15**	**16.84**
S	20 kb	62.04	72.94	69.96
Singletons	20 kb	34.21	42.39	39.04
Number of Mutations	20 kb	62.18	73.12	70.13
Number of Haplotypes	20 kb	17.32	18.08	17.82
Haplotype Diversity	20 kb	0.96	0.98	0.98
Wall’s B	**20 kb**	**0.30**	**0.34**	**0.34**
Wall’s Q	20 kb	0.43	0.47	0.46
Theta W	20 kb	0.00088	0.00104	0.00099
Theta Pi	20 kb	0.00066	0.00078	0.00074
TajD	20 kb	-1.08	-0.91	-1.01
Fu and Li’s D*	20 kb	-1.64	-1.41	-1.51
Fu and Li’s F*	20 kb	-1.71	-1.47	-1.58
Hudson’s Ĉ (rho)	**20 kb**	**16.29**	**82.13**	**15.23**

Table S3 (cont’d). Results for sequence windows of 10 kb (5 kb to each side) and 20 kb (10 kb to each side). These analyses are identical to those presented above, except that a given site was included in the analysis only if all 20 strains showed coverage at that site.

Coding Sequences (remove introns)	All genes (n=11479[a])	Candidate Genes (n=85[a])	P-value[b]
Mean D	-0.69	-0.82	0.12
Median D	-0.87	-1.11	0.10
Mean D*	**-1.03**	**-1.32**	**0.02**
Median D*	-1.20	-1.44	0.09
Mean F*	**-1.06**	**-1.31**	**0.05**
Median F*	-1.17	-1.46	0.11

[a] Removing ‘NA’ values for which metric is undefined (no polymorphism)

[b] Proportion of randomly generated gene sets of the same size that are more extreme than the observed (candidate) gene set.

Table S4. Comparison of Site Frequency Spectrum for Candidate versus Non-Candidate Genes. Mean or median Tajima’s D, Fu and Li’s D*, or Fu and Li’s F*, comparing candidate genes to the genome-wide average.

Metric	5th Percentile for Random Gene Datasets	95th Percentile, Random Gene Datasets	Observed Value for Candidate Genes
Haplotype Diversity	**0.099**	**0.149**	**0.082 (low)**
Wall's B	0.059	0.095	0.080
Wall's Q	0.080	0.113	0.094
Theta W	2.32E-06	1.38E-05	9.56E-06
Theta Pi	1.50E-06	1.25E-05	6.64E-06
Tajima's D	0.71	1.17	1.07
Fu and Li's D*	1.28	1.88	1.79
Fu and Li's F*	1.42	2.10	2.03
Hudson's $\hat{C}$ (rho)	**6.62E+06**	**1.54E+07**	**3.43E+06 (low)**
Number of Segregating Sites[a]	2.67E-05	0.000143838	7.59E-05
Number of Haplotypes[a]	**1.36E-05**	**5.69E-05**	**1.13E-05 (low)**
Number of Singletons[a]	1.09E-05	7.68E-05	5.80E-05

[a] Per site (divided by gene length).

Table S5. Tests of elevated variance in candidate genes compared to random genes. For each evolutionary metric, we tested whether the variance was lower or higher for candidate genes compared to 10,000 datasets consisting of genes chosen at random.

Statistic	Number (and Percentage) of Candidate Genes in the lower 5th percentile of the genomewide distribution[a]	Number (and Percentage) of Candidate Genes in the upper 5th percentile of the genomewide distribution[a]	P-value (two-sided)[b]
Haplotype diversity	0 (0)	7 (7.9)	0.21
Wall's B	0 (0)	5 (5.9)	0.61
Wall's Q	0 (0)	0 (0)	1.00
Theta W	0 (0)	7 (7.9)	0.22
Theta Pi	0 (0)	7 (7.9)	0.22
Tajima's D	**9 (10.6)**	5 (5.9)	**0.04**
Fu and Li's D*	8 (9.4)	6 (7.1)	0.07
Fu and Li's F*	8 (9.4)	4 (4.7)	0.08
Hudson's Ĉ (rho)	0 (0)	0 (0)	1.00
Segregating Sites[c]	0 (0)	5 (5.6)	0.80
Haplotype Number[c]	7 (7.9)	2 (2.2)	0.33
Singletons[3]	0 (0)	4 (4.5)	1.00
p_N	0 (0)	5 (5.7)	0.63
p_S	0 (0)	3 (3.4)	0.80
Omega (p_N/p_S)	0 (0)	**14 (15.2)**	**0.0002**
Number of times $p_N > p_S$[d]	0 (0)	8 (9)	0.08
MK-ratio[e]	0 (0)	2 (4.5)	1.00

[a] Null expectation is that 5% of candidate genes will reside in the top 5% of the genome-wide distribution.

[b] P-value is the result of a two-tailed Fisher's Exact test that compares the number of extreme genes versus not for candidate versus non-candidate genes.

[c] Scaled to gene length

[d] Out of 210 possible pairwise comparisons of 21 sequences (n[n-1]/2 = 21*20/2 = 210). This metric is similar to p_N/p_S but will not result in division by zero when p_S is zero.

[e] MK-ratio = (D_N/D_S)/(P_N/P_S). A value of '1' was added to all values in the 2x2 matrix to prevent division by zero.

Table S6. Number (and Percent) of Candidate Genes that are "Extreme" (in the lower 5th or upper 95th percentile of the genome-wide distribution). Boldface indicates metrics where candidate genes are significantly overrepresented in the tails of the genomewide distribution.

Supplemental Experimental Procedures

Candidate Genes. The candidate genes are described in greater detail in Santorelli et al. 2008 [S2]. Briefly, they consist of approximately 167 insertion sites, 61% of which are insertions into genes, whereas the remaining 39% occur outside of known genes. Seven of these mutants were dropped because we were unable to map the insertion site unambiguously to a single location in the current reference genome, resulting in a total of 160 candidate loci for our analyses. Assays for cheating behaviors on a subset of these mutants showed that ~80% of the mutants arising from the screen cheat the wild-type AX4 strain in head-to-head competition for spore production. All cheater mutants were “facultative” (as opposed to “obligate”) cheaters, in that they were capable of forming fruiting bodies when developed clonally, although they can vary in the total number of spores they produce.

For molecular evolution statistics calculated on genes, candidate genes were the subset of candidate loci where the mutation generating the cheating behavior occurred within a protein-coding gene (n=94). Previous analyses indicate that there are few distinguishing features of these genes compared to other genes in the genome, other than their involvement in cheating behaviors. Candidate genes are present on all six chromosomes. They did not differ from the rest of the genome in their GC content or patterns of codon usage, and they showed no overrepresentation of recognizable protein domains [S2]. Gene ontology (GO) annotations that were significantly enriched were generally involved in protein or amino-acid metabolism, protein modification (e.g., ubiquitination), or signal transduction, but there was no clear process or function that seemed uniquely targeted [S2]. Approximately 61% of candidate loci involved insertions into protein-coding regions, similar to the estimated 62% of the genome that is protein-coding. However, candidate genes were significantly larger than expected by chance (mean size = 2887 bp versus 1662 bp; P<0.001).

Library Preparation.

454. Five micrograms of DNA were sheared by nebulization and fractionated on an agarose gel to isolate 450–550 base fragments. These were used to construct a single-stranded library that was used as template for single-molecule PCR on 28-mm diameter beads in emulsions. The amplified template beads were recovered after emulsion breaking and selective enrichment. Sequencing primer was annealed to the template and the beads were incubated with Bst DNA polymerase, apyrase and single-stranded binding protein. A slurry of the template beads, enzyme beads (required for signal transduction) and packing beads (for Bst DNA polymerase retention) was loaded into the wells of a picotiter plate. The picotiter plate was inserted in the flow cell and subjected to pyro-sequencing on the Genome Sequencer FLX instrument (Roche). The Genome Sequencer FLX flows 100 cycles of four solutions containing either dTTP, aSdATP, dCTP and dGTP reagents, in that order, over the cell. For each dNTP flow, a single 38-s image was captured by a CCD (charge-coupled device) camera on the sequencer. The images were processed in real time to identify template-containing wells and to compute associated signal intensities. The images were further processed for chemical and optical cross-talk, phase errors and read quality before base calling was performed for each template bead.

Illumina. High molecular weight double strand genomic DNA samples were constructed into Illumina paired end libraries according to the manufacturer's protocol (Illumina Inc.). Briefly, 5 µg of genomic DNA in a 100-µl volume was sheared into fragments of approximately 300 bp with the Covaris S2 or E210 system (Covaris, Inc. Woburn, MA). Fragments were processed through DNA End-Repair, and A-tailing and fragments were ligated to Illumina PE adapters. Ligated products were size selected on a 2% low-melt agarose gel, and 290-bp to 320-bp DNA fragments were excised and purified from the gel. This size-selected DNA was PCR-amplified with Illumina PE 1.0 and 2.0 primers using 2x Phusion High-Fidelity PCR master mix for 10 rounds of amplification. Agencourt® XP® Beads (Beckman Coulter Genomics, Inc.; Cat. No. A63882) were used to purify the PCR products. Following bead purification, PCR products were quantified using PicoGreen (Life Technologies; Cat. No. P7589) and their size distribution analyzed using the Agilent Bioanalyzer 2100 DNA Chip 7500 (Agilent; Cat. No. 5067-1506). 15 µl of the 10 mM final library was used for Illumina sequencing.

Shotgun DNA libraries were sequenced on Illumina's Genome Analyzer IIx system according to the manufacturer's specifications. Briefly, sequencing libraries were quantified with an Agilent 2100 Bioanalyzer. Cluster generations were performed on an Illumina cluster station. Sequencing was carried out for each library in a separate, single flow cell lane on the Illumina GA II. Sequencing analysis was done using the Illumina analysis pipeline. Sequencing image files were processed to generate base calls and Phred-like base quality scores and to remove low-quality reads.

SNP calling

454. For the two strains sequenced using 454, we mapped the reads to the reference genome using Atlas-SNP [S3]. Briefly, Atlas-SNP is an integrated short-read assembly and mapping pipeline that uses BLAT [S4] to align reads to the reference genome and cross-match (www.phrap.org/phredphrapconsed.html) to identify all mismatches between the reference genome and the sequencing reads. Reads were mapped with the following parameters: maximum substitution (-s) rate of 10% and maximum insertion rate (-g) of 10%. Following assembly, candidate SNPs were filtered according to the following criteria: an adjusted quality score>=30, a minimum of 2 reads showing the SNP, and a minimum of 80% of the reads covering the site showing the SNP.

Illumina. Sequencing reads were mapped to the reference using MAQ (version 0.7.1), with the following parameters: maximum mismatches (-m) of 9 (for 45 bp reads) or 15 (for 75 bp reads), total quality score (-q) of 203 (45 bp reads) or 338 (75 bp reads). Together, these parameters retain reads that show no more than 20% high quality mismatches at the nucleotide level. Duplicate reads, those showing the same start and end points, were removed from the alignments following mapping, as recommended. Candidate SNPs were called according to the following criteria: at least one read showing the SNP from both the positive and negative strand, a maximum mapping quality of 40 and a minimum consensus quality of 20. For both the Illumina and 454 sequences, any site that did not meet the above criteria designating it as a SNP was assigned the nucleotide of the reference genome at that position if at least one read covered the site and that read showed the reference strain nucleotide, or if at least half of the reads covering the site showed the reference nucleotide; otherwise, the site was assigned as unknown ("N").

Genome Assembly of *D. citrinum.* 454 FLX data generated from *D. citrinum* was assembled using Newbler (454 Life Sciences, Branford, CT). We further improved the assembly using ATLAS GapFill (http://www.hgsc.bcm.tmc.edu/content/atlas-gapfill). The resulting contig N50 length was 4,232 bp, and scaffold N50 length 27,919 bp. The sequence is available from Genbank accession number AJWG00000000.1. MAKER [S5] was used to generate consensus gene predictions derived from *ab initio* models, transcriptome data, and protein similarity. *Ab initio* predictors Augustus [S6] and SNAP [S7] were trained specifically for *D. citrinum* using three rounds of gene annotation bootstrapping starting with *D. discoideum* trained predictors, then training on the resulting *D. citrinum* maker models. RNA-seq data were aligned to the masked reference genome and assembled into transcript models using a tophat/cufflinks pipeline [S8], which were then used as transcript evidence in the MAKER pipeline. In addition to the transcript sequences, several protein databases were provided to the MAKER pipeline for homology evidence: the proteomes of the previously annotated *D. discoideum* and *D. purpureum*. The *D. citrinum* genes orthologous to *D. disodium* genes were identified using the Inparanoid algorithm [S9] based on best reciprocal blast hits of the amino acid sequences from predicted gene models. We estimated synonymous and nonsynonymous nucleotide substitution rates (d_S, d_N) by the maximum likelihood program *codeml* of PAML4 [S10] based on retro-translated protein sequence alignments from GAP4 [S11] global alignment. This process results in a total of 5,923 orthologous genes between *D. citrinum* and *D. discoideum* - of which, 44 were candidate genes. The percentage of candidate genes with a *D. citrinum* ortholog (42%) was similar to the percentage across the genome as a whole (47%).

Population Genomic Analyses. F_{ST} and the difference in allele frequency between populations were calculated for all segregating sites in the genome using scripts written in Ruby and Python. Analyses of population structure were carried out by comparing differences in allele frequency between the two sites, Texas and Virginia. We limited our analysis to SNPs that were diallelic, present in either Texas or Virginia, and genomic sites where we could ascertain SNP presence/absence for a minimum of six strains per geographic site. Where they occurred, negative F_{ST} values were set to zero, as described in [S12]. SNPs were considered to be "candidate" if they occurred in 5-kb or 10-kb windows and "non-candidate" if they fell outside these regions. Distances were chosen based on analyses of LD, indicating that it reaches baseline levels at distances of approximately 20 kb from a focal site.

Nucleotide diversity (Theta Pi), Tajima's *D*, Fu and Li's *D**, Hudson's *C* (recombination, or rho), haplotype diversity, Fay and Wu's *H*, and haplotype number were determined for all genes in the genome and all sequence windows using the program "compute" (available at molpopgen.org). Levels of non-synonymous (p_N) and synonymous (p_S) diversity were calculated using the program "gestimator", and the McDonald-Kreitman tests were calculated using "MKtest" (both available at molpopgen.org.) All analyses were based on the software version 0.8.0. The MK-ratio (=[D_N/D_S]/[P_N/P_S]) was calculated after first adding 1 to every value in the 2x2 matrix to prevent any division by zero. For analyses incorporating *D. citrinum* sequences, we generated multiple sequence alignments using RevTrans and refined them using Muscle, MACSE, and Geneious [S13-S15]. The significance of each population genetic parameter was determined by generating random data sets containing the same number of genes as our candidate gene set. This resampling process generates the distribution under the null hypothesis that candidate genes do not differ from the rest of the genome for the statistic of interest. All statistical analyses were performed in R.

Molecular Evolution as a Function of the Timing or Level of Expression. Parikh et al. [S1], reported the expression level of every gene in the genome every four hours starting from the onset of starvation. We used their categorization of genes as vegetative, developmental, both (vegetative and developmental), or not expressed. To calculate expression levels for vegetative growth, we used log-transformed 0 hr data, and for expression levels during development, we used the average of the log-transformed values for the 8-24 hr timepoints. For each metric, we compared the observed value to the distribution of values in randomly chosen gene sets of the same size.

Supplemental References

S1. Parikh, A., Miranda, E. R., Katoh-Kurasawa, M., Fuller, D., Rot, G., Zagar, L., Curk, T., Sucgang, R., Chen, R., Zupan, B., et al. (2010). Conserved developmental transcriptomes in evolutionarily divergent species. Genome Biol. *11*, R35.

S2. Santorelli, L. A., Thompson, C. R. L., Villegas, E., Svetz, J., Dinh, C., Parikh, A., Sucgang, R., Kuspa, A., Strassmann, J. E., Queller, D. C., et al. (2008). Facultative cheater mutants reveal the genetic complexity of cooperation in social amoebae. Nature *451*, 1107–1110.

S3. Shen, Y., Wan, Z., Coarfa, C., Drabek, R., Chen, L., Ostrowski, E. A., Liu, Y., Weinstock, G. M., Wheeler, D. A., Gibbs, R. A., et al. (2010). A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Res. *20*, 273–280.

S4. Kent, W. J. (2002). BLAT---The BLAST-Like Alignment Tool. Genome Res. *12*, 656–664.

S5. Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A., and Yandell, M. (2007). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. *18*, 188–196.

S6. Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics *19 Suppl 2*, ii215–25.

S7. Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics *5*, 59.

S8. Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

S9. O'Brien, K. P. (2004). Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Research *33*, D476–D480.

S10. Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

S11. Huang X, Brutlag DL (2007). Dynamic use of multiple parameter sets in sequence alignment. (2007). Nucleic Acids Res. *35*, 678–686.

S12. Akey, J. M. (2002). Interrogating a high-density SNP map for signatures of natural selection. Genome Res. *12*, 1805–1814.

S13. Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.

S14. Ranwez, V., Harispe, S., Delsuc, F., and Douzery, E. J. P. (2011). MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PLoS ONE *6*, e22594.

S15. Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics *28*, 1647–1649.