Complete Genome Sequence of the Apicomplexan, *Cryptosporidium parvum*

Mitchell S. Abrahamsen, *et al.*

Science **304**, 441 (2004);
DOI: 10.1126/science.1094786

3. R. Jackendoff, *Foundations of Language: Brain, Grammar, Evolution* (Oxford Univ. Press, Oxford, 2003).
4. Although for Frege (*1*), reference was established relative to objects in the world, here we follow Jackendoff's suggestion (*3*) that this is done relative to objects and the state of affairs as mentally represented.
5. S. Zola-Morgan, L. R. Squire, in *The Development and Neural Bases of Higher Cognitive Functions* (New York Academy of Sciences, New York, 1990), pp. 434–456.
6. N. Chomsky, *Reflections on Language* (Pantheon, New York, 1975).
7. J. Katz, *Semantic Theory* (Harper & Row, New York, 1972).
8. D. Sperber, D. Wilson, *Relevance* (Harvard Univ. Press, Cambridge, MA, 1986).
9. K. I. Forster, in *Sentence Processing*, W. E. Cooper, C. T. Walker, Eds. (Erlbaum, Hillsdale, NJ, 1989), pp. 27–85.
10. H. H. Clark, *Using Language* (Cambridge Univ. Press, Cambridge, 1996).
11. Often word meanings can only be fully determined by invoking world knowledge. For instance, the meaning of "flat" in a "flat road" implies the absence of holes. However, in the expression "a flat tire," it indicates the presence of a hole. The meaning of "finish" in the phrase "Bill finished the book" implies that Bill completed reading the book. However, the phrase "the goat finished the book" can only be interpreted as the goat eating or destroying the book. The examples illustrate that word meaning is often underdetermined and necessarily intertwined with general world knowledge. In such cases, it is hard to see how the integration of lexical meaning and general world knowledge could be strictly separated (*3*, *31*).
12. W. Marslen-Wilson, C. M. Brown, L. K. Tyler, *Lang. Cognit. Process.* **3**, 1 (1988).
13. ERPs for 30 subjects were averaged time-locked to the onset of the critical words, with 40 items per condition. Sentences were presented word by word on the center of a computer screen, with a stimulus onset asynchrony of 600 ms. While subjects were reading the sentences, their EEG was recorded and amplified with a high-cut-off frequency of 70 Hz, a time constant of 8 s, and a sampling frequency of 200 Hz.
14. Materials and methods are available as supporting material on *Science* Online.
15. M. Kutas, S. A. Hillyard, *Science* **207**, 203 (1980).
16. C. Brown, P. Hagoort, *J. Cognit. Neurosci.* **5**, 34 (1993).
17. C. M. Brown, P. Hagoort, in *Architectures and Mechanisms for Language Processing*, M. W. Crocker, M. Pickering, C. Clifton Jr., Eds. (Cambridge Univ. Press, Cambridge, 1999), pp. 213–237.
18. F. Varela *et al.*, *Nature Rev. Neurosci.* **2**, 229 (2001).
19. We obtained TFRs of the single-trial EEG data by convolving complex Morlet wavelets with the EEG data and computing the squared norm for the result of the convolution. We used wavelets with a 7-cycle width, with frequencies ranging from 1 to 70 Hz, in 1-Hz steps. Power values thus obtained were expressed as a percentage change relative to the power in a baseline interval, which was taken from 150 to 0 ms before the onset of the critical word. This was done in order to normalize for individual differences in EEG power and differences in baseline power between different frequency bands. Two relevant time-frequency components were identified: (i) a theta component, ranging from 4 to 7 Hz and from 300 to 800 ms after word onset, and (ii) a gamma component, ranging from 35 to 45 Hz and from 400 to 600 ms after word onset.
20. C. Tallon-Baudry, O. Bertrand, *Trends Cognit. Sci.* **3**, 151 (1999).
21. W. H. R. Miltner *et al.*, *Nature* **397**, 434 (1999).
22. M. Bastiaansen, P. Hagoort, *Cortex* **39** (2003).
23. O. Jensen, C. D. Tesche, *Eur. J. Neurosci.* **15**, 1395 (2002).
24. Whole brain T2*-weighted echo planar imaging blood oxygen level–dependent (EPI-BOLD) fMRI data were acquired with a Siemens Sonata 1.5-T magnetic resonance scanner with interleaved slice ordering, a volume repetition time of 2.48 s, an echo time of 40 ms, a 90° flip angle, 31 horizontal slices, a 64 × 64 slice matrix, and isotropic voxel size of 3.5 × 3.5 × 3.5 mm. For the structural magnetic resonance image, we used a high-resolution (isotropic voxels of 1 mm^3) T1-weighted magnetization-prepared rapid gradient-echo pulse sequence. The fMRI data were preprocessed and analyzed by statistical parametric mapping with SPM99 software (http://www.fil.ion.ucl.ac.uk/spm99).
25. S. E. Petersen *et al.*, *Nature* **331**, 585 (1988).
26. B. T. Gold, R. L. Buckner, *Neuron* **35**, 803 (2002).
27. E. Halgren *et al.*, *J. Psychophysiol.* **88**, 1 (1994).
28. E. Halgren *et al.*, *Neuroimage* **17**, 1101 (2002).
29. M. K. Tanenhaus *et al.*, *Science* **268**, 1632 (1995).
30. J. J. A. van Berkum *et al.*, *J. Cognit. Neurosci.* **11**, 657 (1999).
31. P. A. M. Seuren, *Discourse Semantics* (Basil Blackwell, Oxford, 1985).

32. We thank P. Indefrey, P. Fries, P. A. M. Seuren, and M. van Turennout for helpful discussions. Supported by the Netherlands Organization for Scientific Research, grant no. 400-56-384 (P.H.).

Supporting Online Material
www.sciencemag.org/cgi/content/full/1095455/DC1
Materials and Methods
Fig. S1
References and Notes

8 January 2004; accepted 9 March 2004
Published online 18 March 2004;
10.1126/science.1095455
Include this information when citing this paper.

Complete Genome Sequence of the Apicomplexan, *Cryptosporidium parvum*

Mitchell S. Abrahamsen,[1,2]*† Thomas J. Templeton,[3]† Shinichiro Enomoto,[1] Juan E. Abrahante,[1] Guan Zhu,[4] Cheryl A. Lancto,[1] Mingqi Deng,[1] Chang Liu,[1]‡ Giovanni Widmer,[5] Saul Tzipori,[5] Gregory A. Buck,[6] Ping Xu,[6] Alan T. Bankier,[7] Paul H. Dear,[7] Bernard A. Konfortov,[7] Helen F. Spriggs,[7] Lakshminarayan Iyer,[8] Vivek Anantharaman,[8] L. Aravind,[8] Vivek Kapur[2,9]

The apicomplexan *Cryptosporidium parvum* is an intestinal parasite that affects healthy humans and animals, and causes an unrelenting infection in immunocompromised individuals such as AIDS patients. We report the complete genome sequence of *C. parvum*, type II isolate. Genome analysis identifies extremely streamlined metabolic pathways and a reliance on the host for nutrients. In contrast to *Plasmodium* and *Toxoplasma*, the parasite lacks an apicoplast and its genome, and possesses a degenerate mitochondrion that has lost its genome. Several novel classes of cell-surface and secreted proteins with a potential role in host interactions and pathogenesis were also detected. Elucidation of the core metabolism, including enzymes with high similarities to bacterial and plant counterparts, opens new avenues for drug development.

Cryptosporidium parvum is a globally important intracellular pathogen of humans and animals. The duration of infection and pathogenesis of cryptosporidiosis depends on host immune status, ranging from a severe but self-limiting diarrhea in immunocompetent individuals to a life-threatening, prolonged infection in immunocompromised patients. A substantial degree of morbidity and mortality is associated with infections in AIDS patients. Despite intensive efforts over the past 20 years, there is currently no effective therapy for treating or preventing *C. parvum* infection in humans.

Cryptosporidium belongs to the phylum Apicomplexa, whose members share a common apical secretory apparatus mediating locomotion and tissue or cellular invasion. Many apicomplexans are of medical or veterinary importance, including *Plasmodium*, *Babesia*, *Toxoplasma*, *Neosprora*, *Sarcocystis*, *Cyclospora*, and *Eimeria*. The life cycle of *C. parvum* is similar to that of other cyst-forming apicomplexans (e.g., *Eimeria* and *Toxoplasma*), resulting in the formation of oocysts

[1]Department of Veterinary and Biomedical Science, College of Veterinary Medicine, [2]Biomedical Genomics Center, University of Minnesota, St. Paul, MN 55108, USA. [3]Department of Microbiology and Immunology, Weill Medical College and Program in Immunology, Weill Graduate School of Medical Sciences of Cornell University, New York, NY 10021, USA. [4]Department of Veterinary Pathobiology, College of Veterinary Medicine, Texas A&M University, College Station, TX 77843, USA. [5]Division of Infectious Diseases, Tufts University School of Veterinary Medicine, North Grafton, MA 01536, USA. [6]Center for the Study of Biological Complexity and Department of Microbiology and Immunology, Virginia Commonwealth University, Richmond, VA 23198, USA. [7]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK. [8]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. [9]Department of Microbiology, University of Minnesota, Minneapolis, MN 55455, USA.

*To whom correspondence should be addressed. E-mail: abe@umn.edu
†These authors contributed equally to this work.
‡Present address: Bioinformatics Division, Genetic Research, GlaxoSmithKline Pharmaceuticals, 5 Moore Drive, Research Triangle Park, NC 27009, USA.

that are shed in the feces of infected hosts. *C. parvum* oocysts are highly resistant to environmental stresses, including chlorine treatment of community water supplies; hence, the parasite is an important water- and food-borne pathogen (*1*). The obligate intracellular nature of the parasite's life cycle and the inability to culture the parasite continuously in vitro greatly impair researchers' ability to obtain purified samples of the different developmental stages. The parasite cannot be genetically manipulated, and transformation methodologies are currently unavailable. To begin to address these limitations, we have obtained the complete *C. parvum* genome sequence and its predicted protein complement. (This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the project accession AAEE00000000. The version described in this paper is the first version, AAEE01000000.)

The random shotgun approach was used to obtain the complete DNA sequence (*2*) of the Iowa "type II" isolate of *C. parvum*. This isolate readily transmits disease among numerous mammals, including humans. The resulting genome sequence has roughly 13× genome coverage containing five gaps and 9.1 Mb of total DNA sequence within eight chromosomes. The *C. parvum* genome is thus quite compact relative to the 23-Mb, 14-chromosome genome of *Plasmodium falciparum* (*3*); this size difference is predominantly the result of shorter intergenic regions, fewer introns, and a smaller number of genes (Table 1). Comparison of the assembled sequence of chromosome VI to that of the recently published sequence of chromosome VI (*4*) revealed that our assembly contains an additional 160 kb of sequence and a single gap versus two, with the common sequences displaying a 99.993% sequence identity (*2*).

The relative paucity of introns greatly simplified gene predictions and facilitated annotation (*2*) of predicted open reading frames (ORFs). These analyses provided an estimate of 3807 protein-encoding genes for the *C. parvum* genome, far fewer than the estimated 5300 genes predicted for the *Plasmodium* genome (*3*). This difference is primarily due to the absence of an apicoplast and mitochondrial genome, as well as the presence of fewer genes encoding metabolic functions and variant surface proteins, such as the *P. falciparum* var and rifin molecules (Table 2). An analysis of the encoded protein sequences with the program SEG (*5*) shows that these protein-encoding genes are not enriched in low-complexity sequences (34%) to the extent observed in the proteins from *Plasmodium* (70%).

Our sequence analysis indicates that *Cryptosporidium*, unlike *Plasmodium* and *Toxoplasma*, lacks both mitochondrion and apicoplast genomes. The overall completeness of the genome sequence, together with the fact that similar DNA extraction procedures used to isolate total genomic DNA from *C. parvum* efficiently yielded mitochondrion and apicoplast genomes from *Eimeria* sp. and *Toxoplasma* (*6*, *7*), indicates that the absence of organellar genomes was unlikely to have been the result of methodological error. These conclusions are consistent with the absence of nuclear genes for the DNA replication and translation machinery characteristic of mitochondria and apicoplasts, and with the lack of mitochondrial or apicoplast targeting signals for tRNA synthetases.

A number of putative mitochondrial proteins were identified, including components of a mitochondrial protein import apparatus, chaperones, uncoupling proteins, and solute translocators (table S1). However, the genome does not encode any Krebs cycle enzymes, nor the components constituting the mitochondrial complexes I to IV; this finding indicates that the parasite does not rely on complete oxidation and respiratory chains for synthesizing adenosine triphosphate (ATP). Similar to *Plasmodium*, no orthologs for the γ, δ, or ε subunits or the c subunit of the F_o proton channel were detected (whereas all subunits were found for a V-type ATPase).

Cryptosporidium, like *Eimeria* (*8*) and *Plasmodium*, possesses a pyridine nucleotide transhydrogenase integral membrane protein that may couple reduced nicotinamide adenine dinucleotide (NADH) and reduced nicotinamide adenine dinucleotide phosphate (NADPH) redox to proton translocation across the inner mitochondrial membrane. Unlike *Plasmodium*, the parasite has two copies of the pyridine nucleotide transhydrogenase gene. Also present is a likely mitochondrial membrane–associated, cyanide-resistant alternative oxidase (AOX) that catalyzes the reduction of molecular oxygen by ubiquinol to produce H_2O, but not superoxide or H_2O_2. Several genes were identified as involved in biogenesis of iron-sulfur [Fe-S] complexes with potential mitochondrial targeting signals (e.g., nifS, nifU, frataxin, and ferredoxin), supporting the presence of a limited electron flux in the mitochondrial remnant (table S2).

Our sequence analysis confirms the absence of a plastid genome (*7*) and, additionally, the loss of plastid-associated metabolic pathways including the type II fatty acid synthases (FASs) and isoprenoid synthetic enzymes that

Table 1. General features of the *C. parvum* genome and comparison with other single-celled eukaryotes. Values are derived from respective genome project summaries (*3*, *26*–*28*). ND, not determined.

Feature	*C. parvum*	*P. falciparum*	*S. pombe*	*S. cerevisiae*	*E. cuniculi*
Size (Mbp)	9.1	22.9	12.5	12.5	2.5
(G+C) content (%)	30	19.4	36	38.3	47
No. of genes	3807	5268	4929	5770	1997
Mean gene length (bp) excluding introns	1795	2283	1426	1424	ND
Gene density (bp per gene)	2382	4338	2528	2088	1256
Percent coding	75.3	52.6	57.5	70.5	90
Genes with introns (%)	5	53.9	43	5	ND
Intergenic regions					
(G+C) content %	23.9	13.6	32.4	35.1	45
Mean length (bp)	566	1694	952	515	129
RNAs					
No. of tRNA genes	45	43	174	299	44
No. of 5*S* rRNA genes	6	3	30	100–200	3
No. of 5.8*S*, 18*S*, and 28*S* rRNA units	5	7	200–400	100–200	22

Table 2. Comparison between predicted *C. parvum* and *P. falciparum* proteins.

Feature	*C. parvum*	*P. falciparum**	Common†
Total predicted proteins	3807	5268	1883
Mitochondrial targeted/encoded	17 (0.45%)	246 (4.7%)	15
Apicoplast targeted/encoded	0	581 (11.0%)	0
var/rif/stevor‡	0	236 (4.5%)	0
Annotated as protease§	50 (1.3%)	31 (0.59%)	27
Annotated as transporter‖	69 (1.8%)	34 (0.65%)	34
Assigned EC function¶	167 (4.4%)	389 (7.4%)	113
Hypothetical proteins	925 (24.3%)	3208 (60.9%)	126

*Values indicated for *P. falciparum* are as reported (*3*) with the exception of those for proteins annotated as protease or transporter. †TBLASTN hits ($e < -5$) between *C. parvum* and *P. falciparum*. ‡As reported in (*3*). §Predicted proteins annotated as "protease or peptidase" for *C. parvum* (CryptoGenome database, http://cryptogenome.umn.edu) and *P. falciparum* (PlasmoDB database, http://plasmodb.org). ‖Predicted proteins annotated as "transporter, permease of P-type ATPase" for *C. parvum* (CryptoGenome) and *P. falciparum* (PlasmoDB). ¶Bidirectional BLAST hit ($e < -15$) to orthologs with assigned Enzyme Commission (EC) numbers. Does not include EC assignment numbers for protein kinases or protein phosphatases (due to inconsistent annotation across genomes), or DNA polymerases or RNA polymerases, as a result of issues related to subunit inclusion. (For consistency, 46 proteins were excluded from the reported *P. falciparum* values.)

are otherwise localized to the plastid in other apicomplexans. *C. parvum* fatty acid biosynthesis appears to be cytoplasmic, conducted by a large (8252 amino acids) modular type I FAS (*9*) and possibly by another large enzyme that is related to the multidomain bacterial polyketide synthase (*10*). Comprehensive screening of the *C. parvum* genome sequence also did not detect orthologs of *Plasmodium* nuclear-encoded genes that contain apicoplast-targeting and transit sequences (*11*).

C. parvum metabolism is greatly streamlined relative to that of *Plasmodium*, and in certain ways it is reminiscent of that of another obligate eukaryotic parasite, the microsporidian *Encephalitozoon*. The degeneration of the mitochondrion and associated metabolic capabilities suggests that the parasite largely relies on glycolysis for energy production. The parasite is capable of uptake and catabolism of monosugars (e.g., glucose and fructose) as well as synthesis, storage, and catabolism of polysaccharides such as trehalose and amylopectin. Like many anaerobic organisms, it economizes ATP through the use of pyrophosphate-dependent phosphofructokinases. The conversion of pyruvate to acetyl–coenzyme A (CoA) is catalyzed by an atypical pyruvate-NADPH oxidoreductase (*Cp*PNO) that contains an N-terminal pyruvate–ferredoxin oxidoreductase (PFO) domain fused with a C-terminal NADPH–cytochrome P450 reductase domain (CPR). Such a PFO-CPR fusion has previously been observed only in the euglenozoan protist *Euglena gracilis* (*12*). Acetyl-CoA can be converted to malonyl-CoA, an important precursor for fatty acid and polyketide biosynthesis. Glycolysis leads to several possible organic end products, including lactate, acetate, and ethanol. The production of acetate from acetyl-CoA may be economically beneficial to the parasite via coupling with ATP production.

Ethanol is potentially produced via two independent pathways: (i) from the combination of pyruvate decarboxylase and alcohol dehydrogenase, or (ii) from acetyl-CoA by means of a bifunctional dehydrogenase (*adhE*) with acetaldehyde and alcohol dehydrogenase activities; *adhE* first converts acetyl-CoA to acetaldehyde and then reduces the latter to ethanol. AdhE predominantly occurs in bacteria but has recently been identified in several protozoans, including vertebrate gut parasites such as *Entamoeba* and *Giardia* (*13*, *14*). Adjacent to the *adhE* gene resides a second gene encoding only the AdhE C-terminal Fe-dependent alcohol dehydrogenase domain. This gene product may form a multisubunit complex with AdhE, or it may function as an alternative alcohol dehydrogenase that is specific to certain growth conditions. *C. parvum* has a glycerol 3-phosphate dehydrogenase similar to those of plants, fungi, and the kinetoplastid *Trypanosoma*, but (unlike trypanosomes) the parasite lacks an ortholog of glycerol kinase and thus this pathway does not yield glycerol production. In addition to the modular fatty acid synthase (*Cp*FAS1) and polyketide synthase homolog (*Cp*PKS1), *C. parvum* possesses several fatty acyl–CoA synthases and a fatty acyl elongase that may participate in fatty acid metabolism. Further, enzymes for the metabolism of complex lipids (e.g., glycerolipid and inositol phosphate) were identified in the genome. Fatty acids are apparently not an energy source, because enzymes of the fatty acid oxidative pathway are absent, with the exception of a 3-hydroxyacyl-CoA dehydrogenase.

C. parvum purine metabolism is greatly simplified, retaining only an adenosine kinase and enzymes catalyzing conversions of adenosine 5′-monophosphate (AMP) to inosine, xanthosine, and guanosine 5′-monophosphates (IMP, XMP, and GMP). Among these enzymes, IMP dehydrogenase (IMPDH) is phylogenetically related to ε-proteobacterial IMPDH and is strikingly different from its counterparts in both the host and other apicomplexans (*15*). In contrast to other apicomplexans such as *Toxoplasma gondii* and *P. falciparum*, no gene encoding hypoxanthine-xanthineguanine phosphoribosyltransferase (HXGPRT) is detected, in contrast to a previous report on the activity of this enzyme in *C. parvum* sporozoites (*16*). The absence of HXGPRT suggests that the parasite may rely solely on a single enzyme system including IMPDH to produce GMP from AMP. In contrast to other apicomplexans, the parasite appears to rely on adenosine for purine salvage, a model supported by the identification of an adenosine transporter. Unlike other apicomplexans and many parasitic protists that can synthesize pyrimidines de novo, *C. parvum* relies on pyrimidine salvage and retains the ability for interconversions among uridine and cytidine 5′-monophosphates (UMP and CMP), their deoxy forms (dUMP and dCMP), and dAMP, as well as their corresponding di- and triphosphonucleotides. The parasite has also largely shed the ability to synthesize amino acids de novo, although it retains the ability to convert select amino acids, and instead appears to rely on amino acid uptake from the host by means of a set of at least 11 amino acid transporters (table S2).

Most of the *Cryptosporidium* core processes involved in DNA replication, repair, transcription, and translation conform to the basic eukaryotic blueprint (*2*). The transcriptional apparatus resembles *Plasmodium* in terms of basal transcription machinery. However, a striking numerical difference is seen in the complements of two RNA binding domains, Sm and RRM, between *P. falciparum* (17 and 71 domains, respectively) and *C. parvum* (9 and 51 domains). This reduction results in part from the loss of conserved proteins belonging to the spliceosomal machinery, including all genes encoding Sm domain proteins belonging to the U6 spliceosomal particle, which suggests that this particle activity is degenerate or entirely lost. This reduction in spliceosomal machinery is consistent with the reduced number of predicted introns in *Cryptosporidium* (5%) relative to *Plasmodium* (> 50%). In addition, key components of the small RNA–mediated posttranscriptional gene silencing system are missing, such as the RNA-dependent RNA polymerase, Argonaute, and Dicer orthologs; hence, RNA interference–related technologies are unlikely to be of much value in targeted disruption of genes in *C. parvum*.

Cryptosporidium invasion of columnar brush border epithelial cells has been described as "intracellular, but extracytoplasmic," as the parasite resides on the surface of the intestinal epithelium but lies underneath the host cell membrane. This niche may allow the parasite to evade immune surveillance but take advantage of solute transport across the host microvillus membrane or the extensively convoluted parasitophorous vacuole. Indeed, *Cryptosporidium* has numerous genes (table S2) encoding families of putative sugar transporters (up to 9 genes) and amino acid transporters (11 genes). This is in stark contrast to *Plasmodium*, which has fewer sugar transporters and only one putative amino acid transporter (GenBank identification number 23612372).

As a first step toward identification of multi–drug-resistant pumps, the genome sequence was analyzed for all occurrences of genes encoding multitransmembrane proteins. Notable are a set of four paralogous proteins that belong to the sbmA family (table S2) that are involved in the transport of peptide antibiotics in bacteria. A putative ortholog of the *Plasmodium* chloroquine resistance–linked gene *Pf*CRT (*17*) was also identified, although the parasite does not possess a food vacuole like the one seen in *Plasmodium*.

Unlike *Plasmodium*, *C. parvum* does not possess extensive subtelomeric clusters of antigenically variant proteins (exemplified by the large families of *var* and *rif/stevor* genes) that are involved in immune evasion. In contrast, more than 20 genes were identified that encode mucin-like proteins (*18*, *19*) having hallmarks of extensive Thr or Ser stretches suggestive of glycosylation and signal peptide sequences suggesting secretion (table S2). One notable example is an 11,700–amino acid protein with an uninterrupted stretch of 308 Thr residues (cgd3_720). Although large families of secreted proteins analogous to the *Plasmodium* multigene families were not found, several smaller multigene clusters were observed that encode predicted secreted proteins, with no detectable similarity to proteins from other organisms (Fig. 1, A and B). Within this group, at least four distinct families appear to have emerged through gene expansions specific to the *Cryp-*

tosporidium clade. These families—SKSR, MEDLE, WYLE, FGLN, and GGC—were named after well-conserved sequence motifs (table S2). Reverse transcription polymerase chain reaction (RT-PCR) expression analysis (*20*) of one cluster, a locus of seven adjacent *CpLSP* genes (Fig. 1B), shows coexpression during the course of in vitro development (Fig. 1C).

An additional eight genes were identified that encode proteins having a periodic cysteine structure similar to the *Cryptosporidium* oocyst wall protein; these eight genes are similarly expressed during the onset of oocyst formation and likely participate in the formation of the coccidian rigid oocyst wall in both *Cryptosporidium* and *Toxoplasma* (*21*). Whereas the extracellular proteins described above are of apparent apicomplexan or lineage-specific invention, *Cryptosporidium* possesses many genes encoding secreted proteins having lineage-specific multidomain architectures composed of animal- and bacterial-like extracellular adhesive domains (fig. S1).

Lineage-specific expansions were observed for several proteases (table S2), including an aspartyl protease (six genes), a subtilisin-like protease, a cryptopain-like cysteine protease (five genes), and a *Plasmodium* falcilysin-like (insulin degrading enzyme–like) protease (19 genes). Nine of the *Cryptosporidium* falcilysin genes lack the Zn-chelating "HXXEH" active site motif and are likely to be catalytically inactive copies that may have been reused for specific protein-protein interactions on the cell surface. In contrast to the *Plasmodium* falcilysin, the *Cryptosporidium* genes possess signal peptide sequences and are likely trafficked to a secretory pathway. The expansion of this family suggests either that the proteins have distinct cleavage specificities or that their diversity may be related to evasion of a host immune response.

Completion of the *C. parvum* genome sequence has highlighted the lack of conventional drug targets currently pursued for the control and treatment of other parasitic protists. On the basis of molecular and biochemical studies and drug screening of other apicomplexans, several putative *Cryptosporidium* metabolic pathways or enzymes have been erroneously proposed to be potential drug targets (*22*), including the apicoplast and its associated metabolic pathways, the shikimate pathway, the mannitol cycle, the electron transport chain, and HXGPRT. Nonetheless, complete genome sequence analysis identifies a number of classic and novel molecular candidates for drug exploration, including numerous plant-like and bacterial-like enzymes (tables S3 and S4).

Although the *C. parvum* genome lacks HXGPRT, a potent drug target in other apicomplexans, it has only the single pathway dependent on IMPDH to convert AMP to GMP. The bacterial-type IMPDH may be a promising target because it differs substantially from that of eukaryotic enzymes (*15*). Because of the lack of de novo biosynthetic capacity for purines, pyrimidines, and amino acids, *C. parvum* relies solely on scavenge from the host via a series of transporters, which may be exploited for chemotherapy. *C. parvum* possesses a bacterial-type thymidine kinase, and the role of this enzyme in pyrimidine metabolism and its drug target candidacy should be pursued. The presence of an alternative oxidase, likely targeted to the remnant mitochondrion, gives promise to the study of salicylhydroxamic acid (SHAM), ascofuranone, and their analogs as inhibitors of energy metabolism in the parasite (*23*).

Cryptosporidium possesses at least 15 "plant-like" enzymes that are either absent in or highly divergent from those typically found in mammals (table S3). Within the glycolytic pathway, the plant-like PPi-PFK has been shown to be a potential target in other parasites including *T. gondii*, and PEPCL and PGI appear to be plant-type enzymes in *C. parvum*. Another example is a trehalose-6-phosphate synthase/phosphatase catalyzing trehalose biosynthesis from glucose-6-phosphate and uridine diphosphate–glucose. Trehalose may serve as a sugar storage source or may function as an antidesiccant, antioxidant, or protein stability agent in oocysts, playing a role similar to that of mannitol in *Eimeria* oocysts (*24*). Orthologs of putative *Eimeria* mannitol synthesis enzymes were not found. However, two oxidoreductases (table S2) were identified in *C. parvum*, one of which belongs to the same families as the plant mannose dehydrogenases (*25*) and the other to the plant cinnamyl alcohol dehydrogenases. In principle, these enzymes could synthesize protective polyol compounds, and the former enzyme could use host-derived mannose to synthesize mannitol.

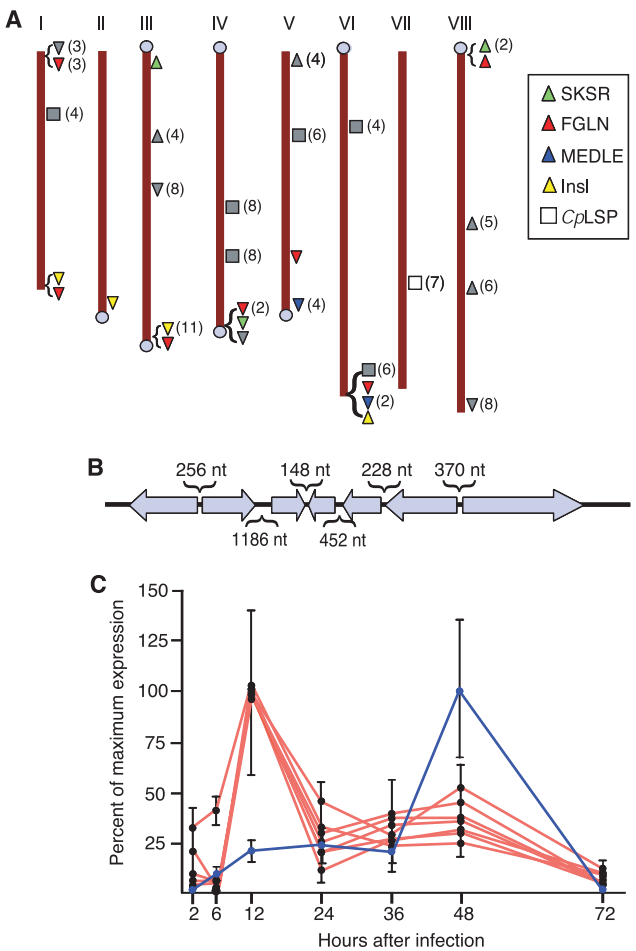

Fig. 1. (**A**) Schematic showing the chromosomal locations of clusters of potentially secreted proteins. Numbers of adjacent genes are indicated in parentheses. Arrows indicate direction of clusters containing unidirectional genes (encoded on the same strand); squares indicate clusters containing genes encoded on both strands. Nonparalogous genes are indicated by solid gray squares or directional triangles; SKSR (green triangles), FGLN (red triangles), and MEDLE (blue triangles) indicate three *C. parvum*–specific families of paralogous genes predominantly located at telomeres. InsI (yellow triangles) indicates an insulinase/falcilysin-like paralogous gene family. *Cp*LSP (white square) indicates the location of a cluster of adjacent large secreted proteins (table S2) that are cotranscriptionally regulated. Identified anchored telomeric repeat sequences are indicated by circles. (**B**) Schematic showing a select locus containing a cluster of coexpressed large secreted proteins (*Cp*LSP). Genes and intergenic regions (regions between identified genes) are drawn to scale at the nucleotide level. The length of the intergenic regions is indicated above or below the locus. (**C**) Relative expression levels of *CpLSP* (red lines) and, as a control, *C. parvum* Hedgehog-type HINT domain gene (blue line) during in vitro development, as determined by semiquantitative RT-PCR using gene-specific primers corresponding to the seven adjacent genes within the *CpLSP* locus as shown in (B). Expression levels from three independent time-course experiments are represented as the ratio of the expression of each gene to that of *C. parvum* 18*S* rRNA present in each of the infected samples (*20*).

References and Notes

1. D. G. Korich *et al.*, *Appl. Environ. Microbiol.* **56**, 1423 (1990).
2. See supporting data on *Science* Online.
3. M. J. Gardner *et al.*, *Nature* **419**, 498 (2002).
4. A. T. Bankier *et al.*, *Genome Res.* **13**, 1787 (2003).
5. J. C. Wootton, *Comput. Chem.* **18**, 269 (1994).

6. X. Cai, A. L. Fuller, L. R. McDougald, G. Zhu, *Gene* **321**, 39 (2003).
7. G. Zhu, M. J. Marchewka, J. S. Keithly, *Microbiology* **146**, 315 (2000).
8. R. A. Kramer *et al*., *Mol. Biochem. Parasitol.* **60**, 327 (1993).
9. G. Zhu *et al*., *Mol. Biochem. Parasitol.* **105**, 253 (2000).
10. G. Zhu *et al*., *Gene* **298**, 79 (2002).
11. B. J. Foth *et al*., *Science* **299**, 705 (2003).
12. C. Rotte, F. Stejskal, G. Zhu, J. S. Keithly, W. Martin, *Mol. Biol. Evol.* **18**, 710 (2001).
13. W. Yang, E. Li, T. Kairong, S. L. Stanley Jr., *Mol. Biochem. Parasitol.* **64**, 253 (1994).
14. B. Rosenthal *et al*., *J. Bacteriol.* **179**, 3736 (1997).
15. B. Striepen *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **99**, 6304 (2002).
16. P. S. Doyle, J. Kanaani, C. C. Wang, *Exp. Parasitol.* **89**, 9 (1998).
17. D. A. Fidock *et al*., *Mol. Cell* **6**, 861 (2000).
18. D. A. Barnes *et al*., *Mol. Biochem. Parasitol.* **96**, 93 (1998).
19. A. M. Cevallos *et al*., *Infect. Immun.* **68**, 5167 (2000).
20. M. S. Abrahamsen, A. A. Shroeder, *Mol. Biochem. Parasitol.* **104**, 141 (1999).
21. T. J. Templeton *et al*., *Infect. Immun.* **72**, 980 (2004).
22. G. H. Coombs, *Parasitol. Today* **15**, 333 (1999).
23. C. Nihei, Y. Fukai, K. Kita, *Biochim. Biophys. Acta* **1587**, 234 (2002).
24. D. M. Schmatz, *Parasitology* **114** (suppl.), S81 (1997).
25. J. D. Williamson *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **92**, 7148 (1995).
26. V. Wood *et al*., *Nature* **415**, 871 (2002).
27. V. Wood *et al*., *Comp. Funct. Genom.* **2**, 143 (2001).
28. M. D. Katinka *et al*., *Nature* **414**, 450 (2001).
29. All DNA sequencing was completed at the Advance Genetics Analysis Center, University of Minnesota.

We thank S. Henning, S. Bertolino, T. Rowan, and H.-W. Chen for their technical assistance, and M. Gottlieb (National Institute of Allergy and Infectious Diseases) for his continual support and encouragement. Supported by NIH grant U01 AI 46397 (M.S.A.).

Supporting Online Material
www.sciencemag.org/cgi/content/full/1094786/DC1
Materials and Methods
SOM Text
Fig. S1
Tables S1 to S4
References

17 December 2003; accepted 15 March 2004
Published online 25 March 2004;
10.1126/science.1094786
Include this information when citing this paper.

Genetic Dissection of Complex Traits with Chromosome Substitution Strains of Mice

Jonathan B. Singer,[1,2]* Annie E. Hill,[3]* Lindsay C. Burrage,[3,4] Keith R. Olszens,[3] Junghan Song,[5]† Monica Justice,[5] William E. O'Brien,[5] David V. Conti,[6]‡ John S. Witte,[6] Eric S. Lander,[1,2,7]§‖ Joseph H. Nadeau[3,4,8]§‖

Chromosome substitution strains (CSSs) have been proposed as a simple and powerful way to identify quantitative trait loci (QTLs) affecting developmental, physiological, and behavioral processes. Here, we report the construction of a complete CSS panel for a vertebrate species. The CSS panel consists of 22 mouse strains, each of which carries a single chromosome substituted from a donor strain (A/J) onto a common host background (C57BL/6J). A survey of 53 traits revealed evidence for 150 QTLs affecting serum levels of sterols and amino acids, diet-induced obesity, and anxiety. These results demonstrate that CSSs greatly facilitate the detection and identification of genes that control the wide diversity of naturally occurring phenotypic variation in the A/J and C57BL/6J inbred strains.

Most traits show substantial genetic variation in natural populations and among inbred strains, reflecting the segregation of quantitative trait loci (QTLs). Genetic identification of QTLs can provide insights into molecular mechanisms of development and physiology, but these studies remain tedious and time consuming (*1*).

The traditional approach for QTL analysis includes two challenging steps. The first step requires arranging a large cross between at least two strains in which hundreds or thousands of progeny are assayed for relevant phenotypes and genotyped for polymorphic markers spanning the genome (*2*, *3*). Because such crosses involve the simultaneous segregation of multiple QTLs, the resulting "phenotypic noise" limits both the power to detect individual QTLs to those with large effect and the precision to localize QTLs to large chromosomal regions. The second step involves molecular identification of the genetic variants that are responsible for each QTL. This step typically requires studying individual QTLs in isolation by performing 5 to 10 generations of backcrosses to construct congenic strains with chromosomal segments carrying alternative alleles of the QTL on an otherwise isogenic background and then interbreeding the congenic strains to carry out fine-structure genetic mapping and cloning. Detection of QTLs and their molecular identification have proven to be serious bottlenecks in studies of complex traits (*1*, *4*, *5*).

We recently proposed (*6*) an approach for QTL analysis that involves prior construction of a panel of chromosome substitution strains (CSSs) between a donor strain (A) and a host strain (B). Strain CSS-*i* carries both copies of chromosome *i* from the donor strain, but all other chromosomes from the host strain are intact and homozygous. A CSS panel partitions the variation between two strains and provides a permanent resource for studying the genetic control of phenotypic variation. Investigators can test individuals from each CSS for any phenotype of interest and immediately infer that a phenotypic difference between the CSS and the host strain implies that at least one QTL resides on the substituted chromosome. Fine-structure mapping of the QTL can then be performed with crosses between the CSS and the host parental strain (*7*, *8*) or with a panel of congenic strains derived directly from the CSS (*9*).

Construction of CSS panels involves successive backcrosses to the host strain, in which progeny carrying a nonrecombinant copy of the desired chromosome are identified in each generation and used as parents for the next backcross to eventually produce progeny heterosomic (A/B) for the desired chromosome on an otherwise host (B/B) background (*6*). These progeny are then intercrossed to produce progeny homosomic (A/A) for the desired chromosome. Although the concept is straightforward, CSS construction has only become feasible with the availability of complete genetic maps that can be used to trace inheritance throughout the genome. [The exception is *Drosophila melanogaster,* in which special balancer chromosomes that suppress recombination can be used for chromosome substitution (*10*)].

[1]The Broad Institute, Cambridge, MA 02142, USA. [2]Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA. [3]Department of Genetics, Case Western Reserve University School of Medicine, Cleveland, OH 44106, USA. [4]Center for Computational Genomics and Systems Biology, Case Western Reserve University, Cleveland, OH 44106, USA. [5]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA. [6]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH 44106, USA. [7]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [8]Center for Human Genetics, University Hospitals of Cleveland, Cleveland, OH 44106, USA.

*These authors contributed equally to this work.
†Present address: Department of Laboratory Medicine, Seoul National University Bundang Hospital 300, Kumi-dong, Bundang-ku, Sungnam, Kyungki 463-707, Korea.
‡Present address: Department of Preventive Medicine, University of Southern California, Los Angeles, CA 90089, USA.
§These authors co-supervised this work.
‖To whom correspondence should be addressed. E-mail: jhn4@cwru.edu (J.H.N.); lander@broad.mit.edu (E.S.L.)