expression of *Sxl* (ref. 28). Moreover, using a very sensitive test[29], we determined that infection does not alter the effectiveness of the primary sex-determination signal (data not shown), perturbations of which can cause sex-specific lethality owing to inappropriate somatic expression of *Sxl*. Nevertheless, the possibility should be explored that *Wolbachia*-induced male killing reported for other *Drosophila* species[6] may be caused by inappropriate activation of *Sxl*.

Although it may seem surprising that infection with a parasite would reverse the deleterious effect of a mutation in the host genome, particularly when the isolation of that mutation had nothing to do with infection, such surprise should be tempered by the fact that the interaction described here between host and parasite mimics a naturally occurring situation mentioned above that was reported recently for the parasitic wasp *Asobara tabida*[9]. Moreover, in light of the fact that *Wolbachia* is a parasite that is known to manipulate host reproductive and sex-determination systems, it does not seem unreasonable that the host gene with which it interacts in *Drosophila* is the master regulator of sex-determination and a gene essential for oogenesis. The fact that the interacting gene in this case has been studied so extensively and belongs to a model experimental organism can be exploited to yield further insights into the mechanism by which this parasite takes advantage of its various arthropod hosts. □

Received 7 February; accepted 23 April 2002; doi:10.1038/nature00843.

1. Werren, J. H. Biology of *Wolbachia*. *Annu. Rev. Entomol.* **42,** 587–609 (1997).
2. Knight, J. Meet the Herod bug. *Nature* **412,** 12–14 (2001).
3. Werren, J. H. & O'Neill, S. L. in *Influential Passengers: Inherited Microorganisms and Arthropod Reproduction* (eds O'Neill, S. L., Hoffman, A. A. & Werren, J. H.) 1–41 (Oxford Univ. Press, Oxford, 1997).
4. Huigens, M. E. *et al.* Infectious parthenogenesis. *Nature* **405,** 178–179 (2000).
5. Bouchon, D., Rigaud, T. & Juchault, P. Evidence for widespread *Wolbachia* infection in isopod crustaceans: molecular identification and host feminization. *Proc. R. Soc. Lond. B* **265,** 1081–1090 (1998).
6. Hurst, G. D. D., Johnson, A. P., Schulenburg, J. H. G. & Fuyama, Y. Male-killing *Wolbachia* in *Drosophila*: a temperature-sensitive trait with a threshold bacterial density. *Genetics* **156,** 699–709 (2000).
7. Boyle, L., O'Neill, S. L., Robertson, H. M. & Karr, T. L. Interspecific and intraspecific horizontal transfer of *Wolbachia* in *Drosophila*. *Science* **260,** 1796–1799 (1993).
8. Bordenstein, S. R., O'Hara, F. P. & Werren, J. H. *Wolbachia*-induced incompatibility precedes other hybrid incompatibilities in *Nasonia*. *Nature* **409,** 707–710 (2001).
9. Dedeine, F. *et al.* Removing symbiotic *Wolbachia* bacteria specifically inhibits oogenesis in a parasitic wasp. *Proc. Natl Acad. Sci. USA* **98,** 6247–6252 (2001).
10. Bourtzis, K., Nirgianaki, A., Markakis, G. & Savakis, C. *Wolbachia* infection and cytoplasmic incompatibility in *Drosophila* species. *Genetics* **144,** 1063–1073 (1996).
11. Min, K. T. & Benzer, S. *Wolbachia*, normally a symbiont of *Drosophila*, can be virulent, causing degeneration and early death. *Proc. Natl Acad. Sci. USA* **94,** 10792–10796 (1997).
12. Cline, T. W. & Meyer, B. J. Vive la différence: males vs females in flies vs worms. *Annu. Rev. Genet.* **30,** 637–702 (1996).
13. Schupbach, T. Normal female germ cell differentiation requires the female X-chromosome to autosome ratio and expression of *Sex-lethal* in *Drosophila melanogaster*. *Genetics* **109,** 529–548 (1985).
14. Cook, K. R. *Regulation of Recombination and Oogenesis by the* ovarian tumor, Sex-lethal, *and* ovo *Genes of* Drosophila melanogaster. Thesis no. 381, Univ. Iowa (1993).
15. Salz, H. K., Cline, T. W. & Schedl, P. Functional changes associated with structural alterations induced by mobilization of a *P* element inserted in the *Sex-lethal* gene of *Drosophila*. *Genetics* **117,** 221–231 (1987).
16. Perrimon, N., Mohler, D., Engstrom, L. & Mahowald, A. P. X-linked female-sterile loci in *Drosophila melanogaster*. *Genetics* **113,** 695–712 (1986).
17. Bopp, D., Horabin, J. I., Lersch, R. A., Cline, T. W. & Schedl, P. Expression of the *Sex-lethal* gene is controlled at multiple levels during *Drosophila* oogenesis. *Development* **118,** 797–812 (1993).
18. Dines, J. L. *New Aspects of Functional Complexity for the Master Regulator of* Drosophila melanogaster *Sex Determination* Thesis no. 319, Univ. California, Berkeley (2001).
19. O'Neill, S. L., Giordano, R., Colbert, A. M. E., Karr, T. L. & Robertson, H. M. 16S rRNA phylogenetic analysis of the bacterial endosymbionts associated with cytoplasmic incompatibility in insects. *Proc. Natl Acad. Sci. USA* **89,** 2699–2702 (1999).
20. Bopp, D., Schutt, C., Puro, J., Huang, H. & Nothiger, R. Recombination and disjunction in female germ cells of *Drosophila* depend on the germline activity of the gene *Sex-lethal*. *Development* **126,** 5785–5794 (1999).
21. Dines, J., Lersch, B., Lu, B., Bell, M. & Cline, T. W. Functional specialization of SEX-LETHAL protein isoforms. *Annu. Drosophila Res. Conf. Program Abs. Vol.* **39,** a245 (1998).
22. Salz, H. K. *et al.* The *Drosophila* female-specific sex-determination gene, *Sex-lethal*, has stage-, tissue-, and sex-specific RNAs suggesting multiple modes of regulation. *Genes Dev.* **3,** 708–719 (1989).
23. Oliver, B., Perrimon, N. & Mahowald, A. P. Genetic evidence that the *sans fille* locus is involved in *Drosophila* sex determination. *Genetics* **120,** 159–172 (1988).
24. Steinmann-Zwicky, M. Sex determination in *Drosophila*: the X-chromosomal gene *liz* is required for *Sxl* activity. *EMBO J.* **7,** 3889–3898 (1988).
25. Pauli, D., Oliver, B. & Mahowald, A. P. The role of the *ovarian tumor* locus in *Drosophila melanogaster* germline sex determination. *Development* **119,** 123–134 (1993).
26. Page, S. L., McKim, K. S., Deneen, B., Van Hook, T. L. & Hawley, S. R. Genetic studies of *mei-P26* reveal a link between the processes that control germ cell proliferation in both sexes and those that control meiotic exchange in *Drosophila*. *Genetics* **155,** 1757–1772 (2000).
27. Hager, J. H. & Cline, T. W. Induction of female *Sex-lethal* RNA splicing in male germ cells: implications for *Drosophila* germline sex determination. *Development* **124,** 5033–5048 (1997).
28. Cline, T. W. A male-specific lethal mutation in *Drosophila melanogaster* that transforms sex. *Dev. Biol.* **72,** 266–275 (1979).
29. Cline, T. W. Evidence that *sisterless-a* and *sisterless-b* are two of several discrete 'numerator elements' of the *X/A* sex determination signal in *Drosophila* that switch *Sxl* between two alternative stable expression states. *Genetics* **119,** 829–862 (1988).

Acknowledgements

We thank L. Sefton for generating the original suppressed Sxl^{f4} strain, D. Presgraves for the *y w* CS *Wolbachia* strain, and B. J. Meyer for comments on the manuscript.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.W.C. (e-mail: sxlcline@uclink.berkeley.edu).

Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*

Gernot Glöckner*, Ludwig Eichinger†, Karol Szafranski*, Justin A. Pachebat‡, Alan T. Bankier‡, Paul H. Dear‡, Rüdiger Lehmann*, Cornelia Baumgart*, Genís Parra§, Josep F. Abril§, Roderic Guigó§, Kai Kumpf*, Budi Tunggal†, the *Dictyostelium* Genome Sequencing Consortium||, Edward Cox¶, Michael A. Quail#, Matthias Platzer*, André Rosenthal☆ & Angelika A. Noegel†

** IMB Jena, Department of Genome Analysis, Beutenbergstr. 11, 07745 Jena, Germany*
† Center for Biochemistry, Medical Faculty, University of Cologne, Joseph-Stelzmann-Str. 52, 50931 Köln, Germany
‡ Medical Research Council Laboratory of Molecular Biology, MRC Centre, Hills Road, Cambridge CB2 2QH, UK
§ Grup de Recerca en Informatica Biomedica, Institut Municipal d'Inveastigació Mèdica, Universitat Pompeu Fabra, Centre de Regulació Genòmica, 08003 Barcelona, Spain
¶ Princeton University, Princeton, New Jersey 08544, USA
 # The Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK
☆ Friedrich Schiller Universität, 07743 Jena, Germany
|| A full list of authors appears at the end of this paper

The genome of the lower eukaryote *Dictyostelium discoideum* comprises six chromosomes. Here we report the sequence of the largest, chromosome 2, which at 8 megabases (Mb) represents about 25% of the genome. Despite an A + T content of nearly 80%, the chromosome codes for 2,799 predicted protein coding genes and 73 transfer RNA genes. This gene density, about 1 gene per 2.6 kilobases (kb), is surpassed only by *Saccharomyces cerevisiae* (one per 2 kb) and is similar to that of *Schizosaccharomyces pombe* (one per 2.5 kb)[1,2]. If we assume that the other chromosomes have a similar gene density, we can expect around 11,000 genes in the *D. discoideum* genome. A significant number of the genes show higher similarities to genes of vertebrates than to those of other fully sequenced eukaryotes[1–6]. This analysis strengthens the view that the evolutionary position of *D. discoideum* is located before the branching of metazoa and fungi but after the divergence of the plant kingdom[7], placing it close to the base of metazoan evolution.

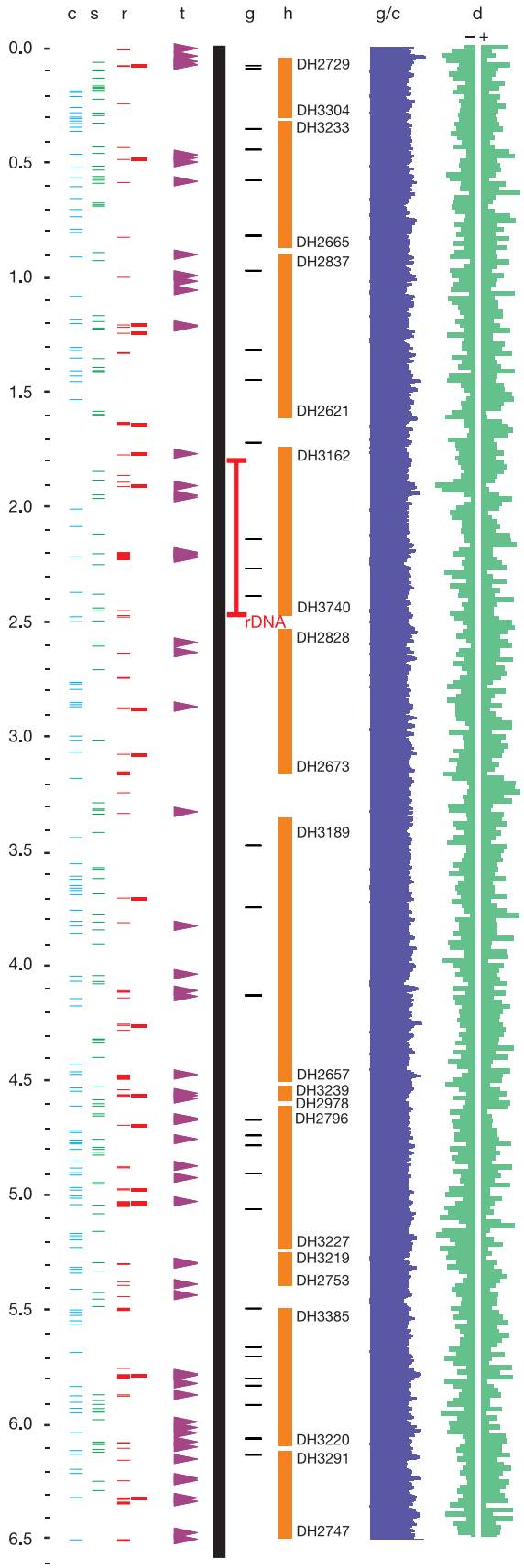

Figure 1 Feature distribution on chromosome 2. Only the linked portion (6.5 Mb) is shown (solid black line). Clone gaps (c), sequence gaps (s), repeat elements (r; heavier bars to their right indicate unresolvable clusters), tRNA genes (t) and genes (g) used to seed assembly are shown. HAPPY linkage groups (h) were used to guide assembly; only the endmost markers in each group are named. G+C content (g/c), strand-specific coding sequence density (d), the ribosomal DNA copy, and the duplicated region above it (represented here as a single copy) are shown. The centromere and telomere are respectively above and below the portion shown. An expanded version is at http://genome.imb-jena.de/dictyostelium/chr2/Chr2map.html.

The natural habitat of *D. discoideum* is deciduous forest soil where the amoeboid cells feed on bacteria by phagocytosis and multiply by equal mitotic division. Exhaustion of the food source triggers a developmental programme, in which more than 100,000 cells aggregate by chemotaxis to form a multicellular structure. Morphogenesis and cell differentiation then culminate in the production of spores, enabling the organism to survive unfavourable conditions[8]. *D. discoideum* therefore lies at the borderline between free-living cells and multicellular organisms, making it ideal for the study of cellular differentiation and integration. Its haploid genome, ease of culture and genetic manipulability make it amenable to biochemical, genetic, and cell-biological approaches[9]. This allows the dissection of the molecular basis of the most fundamental cellular processes: differentiation, signal transduction, phagocytosis, cytokinesis, cell motility and chemotaxis[10–12].

To provide the basis for genome-wide investigations an international effort was initiated[13] to sequence the ~34-Mb genome of *D. discoideum*, strain AX4. Besides six chromosomes ranging from 4 to 8 Mb (refs 14, 15), the nucleus harbours approximately 100 copies of a ~90-kb palindromic chromosome containing the ribosomal RNA genes. The high A+T content (78%, exceeded only by *Plasmodium falciparum* at 80%; refs 16, 17) coupled with a high density of repetitive elements, posed severe challenges for genome sequencing. To reduce the complexity of the assembly task, the genome was analysed chromosome by chromosome, using a whole chromosome shotgun (WCS) approach. The chromosomal libraries were only ~50% pure and contained clones derived from other chromosomes, so we developed an iterative and integrated assembly strategy. This allowed us to identify contiguous DNA sequences (contigs) originating from chromosome 2 and to bridge difficult sequences. Briefly (see Methods), nonrepetitive reads from the chromosome 2-enriched libraries were binned with those from the other WCS projects, and sequences of known chromosome 2 genes were used as 'seeds' around which to build contigs. These were extended using sequence data and supplemented using read-pair information and BLAST (http://blast.wustl@adu/) analysis. To confirm the chromosomal assignment of these contigs we used the relative frequencies of the constituent sequences in the chromosomally enriched libraries of the various WCS projects.

The high A+T content, the existence of many repetitive elements and the fact that clones larger than about 5 kb were unstable in *Escherichia coli*[18,19], precluding the use of large-insert bacterial clones as second-source templates, led to three types of gaps. The first type could not be spanned by plasmid clones ('clone gaps'), presumably owing to the instability of some of the intergenic regions, which have A+T contents of up to 98%. The second type arose from clusters of repetitive elements, which could not be unambiguously resolved ('repeat gaps'). The third type ('sequence gaps') were spanned by clones which, owing to their content of long homopolymer runs (even more abundant and longer than in *P. falciparum*) or lack of targets for custom primers, were recalcitrant to repeated attempts at sequencing. Contigs divided by sequence gaps were linked by read-pair information to produce larger 'scaffolds' with a total size of 7.5 Mb. The majority of these scaffolds were then connected, oriented and their internal structure validated by using mapped genes, circular yeast artificial chromosomes (cYACs) and HAPPY map[20] data. This yielded a 'linked portion' spanning 6.5 Mb of the chromosome (Fig. 1; Table 1; and Supplementary Information; see also http://genome. imb-jena.de/dictyostelium/chr2/Chr2map.html). Although many of the sequence and

Table 1 Features of chromosome 2

Feature	Value
Calculated total length (Mb)*	8.0–8.1
Total length of sequence contigs (Mb)*	7.52
Cumulated length of 71 small orphan unlinked contigs (Mb)	0.4
Number of loci containing complex repetitive elements**	58
Resolved loci	40
Unresolved loci	18
Number of tRNAs	73
Genes*†	
Predicted number	2,799
Density	1 gene/2.6 kb
Average length (bases)	1,626
Number of genes with ESTs	1,120 (40%)
AT content (%)	
Exons	72
Introns	87
Intergenic	86
Whole chromosome	77.8
Exons (coding)	
Number	6,398
Average exon number/gene	2.29
Average size (bases)	711
Introns	
Number	3,587
Average size (bases)	177
Intergenic regions	
Average size (bases)	786
Intronless genes (%)	893 (32)

*Excluding duplication of 0.7 Mb.
**In 6.5-Mb linked portion of chromosome 2.
†Excluding genes coded for in repeat loci.

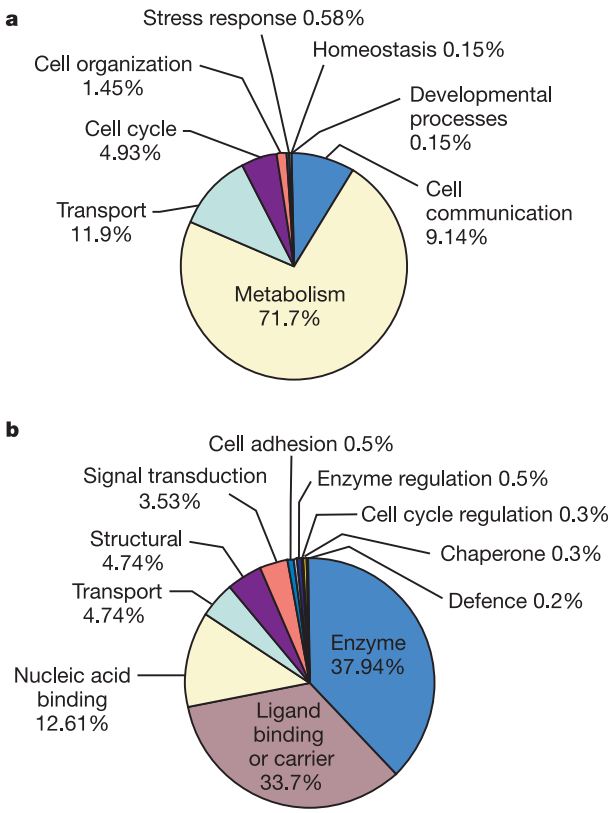

Figure 2 Functional classification of *D. discoideum* chromosome 2-coded proteins. We used the GO terminology (http://whitefly.lbl.gov/annot/go/database/index.html) for the automated classification of proteins in process (**a**) and function (**b**) groups according to their InterPro domains. The process groups contain 689 proteins, the function groups 991 proteins. Proteins with InterPro domains but no GO assignment (424) or proteins without Interpro domains (1,319) were not characterized. Currently no *D. discoideum*-specific GO terms are defined, thus leaving some of the functionally characterized *D. discoideum*-specific genes unclassified.

clone gaps have been closed, those that remain (95 sequence gaps and 89 clone gaps, totalling an estimated 150 kb) appear intractable. The most resistant gaps have been those containing the most A+T-rich DNA, and are hence least likely to contain sequences of biological relevance.

D. discoideum chromosomes have been reported to be acro- or telocentric with the centromere embedded in a large cluster of long terminal repeat retrotransposons (DIRS-1) composed of more than 40 elements[15,19]. The fine structure of this cluster, which lies outside the linked portion of the chromosome shown in Fig. 1, could not be resolved because of the low polymorphism rates of the complex repetitive elements[19]. It spans up to 0.5 Mb and also contains copies from other transposon families and small repetitive elements. Overall, the number of repetitive elements in *D. discoideum* genomic DNA is high compared to *S. cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*. Chromosome 2 harbours all previously described *D. discoideum* complex repetitive elements, mainly organized in clusters of intact and truncated elements[19]. There are 58 such loci (each consisting of one or more such elements) on the linked portion of chromosome 2. The fine structure of 18 of these could not be resolved and remain as 'repeat gaps' (Fig. 1; Table 1). Altogether, we estimate that complex repetitive elements represent 10.2% (approximately 0.8 Mb) of chromosome 2, corresponding well with the estimate of 9.6% for the entire genome[19]. On the basis of the combined sizes of the sequence scaffolds, the clone and sequence gaps, and the unresolved repeat regions (including the pericentromeric region), we calculate the size of the chromosome to be 8.1 Mb.

A duplication of approximately 700 kb is thought to have occurred after the separation of the laboratory strains AX2 and AX4 (ref. 15). We detected an inverse tandem repeat of similar size between the HAPPY Map markers DH3162 and DH3740, bordered at the telomeric end by an almost complete copy of the extrachromosomal rDNA palindrome (Fig. 1). This might represent a chromosomal master copy for the generation of the extrachromosomal rDNA palindrome after sexual recombination, as in *Tetrahymena thermophila*[21]. The second copy of the duplication was excluded from calculations of chromosome length and gene number.

We find that most features of the chromosome (G+C content, coding sequence density (CDS) and complex repetitive elements) are evenly distributed over the 6.5-Mb linked portion, although the distribution of the transfer RNA genes shows a slight bias towards the telomere (Fig. 1; http://genome.imb-jena.de/dictyostelium/chr2/Chr2map.html). We used gene prediction programs and database searches to determine and annotate the 2,799 putative genes of chromosome 2. A further 124 putative genes coded for by complex repetitive elements were excluded from further analyses. *D. discoideum* genes in general have few and small introns, with an average of 1.2 introns per gene. Intron length and distribution is comparable to that of *P. falciparum* and other lower eukaryotes. The mean A+T content in exons is 72%, whereas it is 87% in introns, and 86% in intergenic regions (Table 1). This extreme compositional bias may help to delineate the introns during splicing, as has been suggested in *Arabidopsis thaliana*. In support of this hypothesis, *D. discoideum* introns do contain the canonical GT–AG dinucleotides but, unusually among fully sequenced eukaryotes, all information is confined to the intron side of the splice site[22].

Turning to the gene content, expressed sequence tags (ESTs) exist for 40% (1,120) of the predicted genes (Table 1)[23]. BLAST searches against the protein sets of completely sequenced eukaryotic genomes as well as against SwissProt and TrEMBL databases showed that 45% (1,260) of the putative *D. discoideum* genes had a match ($P < 10^{-15}$), leaving the proportion of unique genes (55%) comparable to that observed for other eukaryotes. About 53% (1,480) of the putative genes contained domains defined in the InterPro database (http://www.ebi.ac.uk/interpro)[24]; again, this proportion is comparable to other eukaryotes[6]. In total, EST, protein, and/or InterPro matches provide support for 1,960 of the 2,799 predicted

Table 2 Most frequent InterPro domains

Domain	Description	DD	SC	AT	CE	DM	HS
IPR001687	ATP/GTP-binding site motif A (P-loop)*	6.07	0.57	0.61	0.32	0.46	0.33
IPR000694	Proline-rich region	3.72	NA	NA	NA	NA	NA
IPR000561	EGF-like domain*	2.18	0.02	0.16	0.68	0.62	1.28
IPR000719	Eukaryotic protein kinase	1.93	1.91	4.07	2.34	1.79	2.64
IPR002290	Serine/threonine protein kinase	1.89	1.83	3.34	1.33	1.22	1.83
IPR001245	Tyrosine protein kinase	1.71	0.05	1.84	0.84	0.65	1.22
IPR001680	G-protein beta WD-40 repeats	1.11	1.63	1.02	0.80	1.31	1.34
IPR003593	AAA ATPase superfamily*	1.11	0.95	0.90	0.40	0.56	0.46
IPR000051	SAM (and some other nucleotide) binding motif	0.89	0.33	0.40	0.25	0.28	0.20
IPR001849	Pleckstrin homology (PH) domain	0.89	0.47	0.12	0.41	0.54	1.24
IPR002048	EF-hand*	0.86	0.26	0.85	0.65	0.93	1.15
IPR001841	RING finger	0.82	0.65	1.82	0.81	0.85	1.20
IPR002085	Zinc-containing alcohol dehydrogenase superfamily	0.82	0.34	0.15	0.06	0.07	0.08
IPR000794	Beta-ketoacyl synthase*	0.79	0.03	0.02	0.02	0.03	0.01
IPR003579	RAS small GTPases, Rab subfamily	0.79	0.15	0.23	0.15	0.21	0.22
IPR001611	Leucine-rich repeat	0.75	0.13	1.93	0.33	0.83	0.74
IPR003577	RAS small GTPases, Ras subfamily	0.75	0.05	0.00	0.06	0.07	0.10
IPR003880	Phosphopantetheine attachment site	0.75	0.10	0.21	0.15	0.28	0.08
IPR000504	RNA-binding region RNP-1 (RNA recognition motif)	0.71	0.93	0.96	0.69	1.13	1.25
IPR000873	AMP-dependent synthetase and ligase	0.71	0.18	0.17	0.17	0.25	0.16
IPR003578	RAS small GTPases, Rho subfamily	0.71	0.10	0.04	0.05	0.04	0.10
IPR000345	Cytochrome *c* family haem-binding site	0.68	0.10	0.58	0.31	0.31	0.37
IPR001227	Acyl transferase domain*	0.68	0.02	0.00	0.02	0.03	0.01
IPR001806	Ras GTPase superfamily	0.68	0.36	0.38	0.33	0.51	0.60
IPR000477	RNA-directed DNA polymerase (Reverse transcriptase)	0.64	0.08	0.50	0.46	0.09	0.14
IPR001064	Zinc-finger GCS-type	0.64	0.10	0.07	0.04	0.07	0.14
IPR002110	Ankyrin-repeat*	0.64	0.29	0.44	0.53	0.62	0.91
IPR001601	Generic methyl-transferase	0.61	0.10	0.22	0.06	0.07	0.06
IPR001410	DEAD/DEAH box helicase	0.57	1.19	0.53	0.44	0.54	0.52
IPR002106	Aminoacyl-transfer RNA synthetases class-II	0.57	0.36	0.35	0.59	0.41	0.20

Occurrence of the thirty most frequent InterPro domains on *D. discoideum* chromosome 2 (DD; including repetitive elements) and fully sequenced eukaryotes. The percentage of genes in each organism that contain the respective domain type is given. SC, *S. cerevisiae*; AT, *A. thaliana*; CE, *C. elegans*; DM, *D. melanogaster*; HS, *H. sapiens*. The data for SC, CE, DM, HS and AT were taken from http://www.ebi.ac.uk/proteome/. NA, not analysed.
*These entries are discussed in the text.

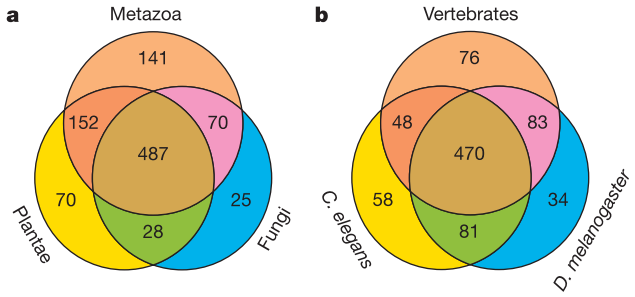

Figure 3 Phylum-specific distribution of proteins. **a**, For comparison with the chromosome 2 translated genes, plants are represented by the full protein set of *A. thaliana* and fungi by *S. cerevisiae* plus *S. pombe*. Metazoans are represented by *D. melanogaster*, *C. elegans*, the (as yet incomplete) *Homo sapiens* protein set, plus annotated nonhuman vertebrate proteins from the SwissProt database. **b**, Distribution of *D. discoideum* genes with $P < 10^{-15}$ between fully sequenced metazoan species. The vertebrate gene set is as in **a**.

genes. Applying the gene ontology (GO) terminology for the automated classification of proteins (http://whitefly.lbl.gov/annot/go/database/index.html) we could attribute functions and/or processes to 1,026 (37%) of the predicted protein products (Fig. 2). Slightly more of these proteins could be classified to functions (991) than to process (689) categories, and 654 out of the 1,026 classified proteins are present in both categories. Forty-seven per cent of the putative proteins remain unclassified and a further 15% of all proteins could not be categorized because their corresponding InterPro domains are not yet assigned to GO terms (Fig. 2).

Because *D. discoideum* undergoes differentiation and development we might expect a significant number of genes associated with multicellular life. In fact, a remarkably high proportion of GO-classified proteins are grouped into the cell communication category (9.14%) and involved in signal transduction or cell adhesion, or comprise cytoskeletal proteins containing signalling domains. As expected, analysis of InterPro matches reveals that domains required for cell motility, signalling, surface attachment and cytoskeletal functions are considerably more abundant than in yeast. When we compare the most frequent InterPro domains of chromosome 2 genes to those of other species (Table 2), the ATP/GTP-binding site motif A (P-loop) is strongly over-represented (6.07% of the predicted genes on the *D. discoideum* chromosome 2 carry this motif, versus 0.33% in human), whereas the epidermal growth factor (EGF)-like domain is over-represented only slightly compared to human (2.18% versus 1.28%), but strongly in comparison to yeast (0.02%). The AAA ATPase superfamily domain is found in comparable proportions in *D. discoideum*, *S. cerevisiae* and *A. thaliana*, but is less abundant in *C. elegans*, *D. melanogaster* and human, whereas the proportions of the Ca^{2+}-binding EF-hand domain and the ankyrin repeat are roughly comparable in all organisms with the exception of yeast, where they are less abundant. We have also identified many beta-ketoacyl synthase and acyl transferase domains, which are hardly present in the other organisms considered here. In *D. discoideum* many of these domains are part of polyketide synthases, which are exceptionally large, multifunctional proteins, primarily present in actinomycetes, bacilli and filamentous fungi. The compounds built via the polyketide synthase pathways might enable *D. discoideum* to defend itself against its natural competitors.

Many *D. discoideum* genes—particularly those involved in signalling and cell movement—are known to be present as multiple copies or as members of large gene families. This is supported by our analysis of chromosome 2, which contains 130 genes present as two or more copies ($P < 10^{-30}$ and sequence similarity over the complete length), amounting to 337 (12%) of the predicted genes. Because paralogues on the other chromosomes have not been taken into account, the number of singletons will further decrease when all chromosomes have been analysed. We have found ten genes for members of the Ras-related small GTP-binding protein family and nine genes sharing the RasGEF domain. Furthermore, we identified another G protein with homology to Gα2, a component of the cyclic AMP signalling system, and also residing on chromosome 2. We have also found two more members of the G-protein coupled receptor family. Surprisingly, these proteins have highest homology to GABA (γ-aminobutyric acid) receptors, which have not yet been found outside the metazoan branch. Genes coding for components of the cytoskeleton are frequently present in multiple copies. Of the ~27 actin genes (including pseudogenes) in the *D. discoideum* genome[19], thirteen are present on chromosome 2 and ten of these translate into identical protein sequences. Chromosome 2 harbours several genes coding for motor proteins, among which are six genes for different unconventional myosins. Although the cytoskeleton has been intensively studied, we have found putative new paralogues for profilin I/II, fimbrin, cofilin 1/2 and the unconventional myosin gene family. The discovery of additional putative paralogues of cytoskeletal proteins supports the concept of functional redundancy in the cytoskeletal system[12]. The ABC (ATP-binding cassette) transporter family is probably one of the largest in the genome. There are thirteen such genes on chromosome 2, including several members of the ABC A subfamily whose occurrence has been restricted to multicellular eukaryotes. ABC transporters use the energy of ATP hydrolysis to translocate specific substrates across cellular membranes. Mutations in many of the human genes coding for ABC transporters are associated with disease such as cystic fibrosis, Stargardt's disease or hyperinsulinism. We have found genes on chromosome 2 with high similarities to the

Table 3 *D. discoideum* chromosome 2 genes with similarity to human disease genes

Disease (gene symbol)	OMIM number	Accession number	*D. discoideum* gene	BLASTP value ($<1.0 \times 10^{-50}$)
Renal tubular acidosis (ATP6B1)	192132	AAD11943	dd_01070	4.0e-246 (0)
Immunodeficiency (DNA Ligase 1)	126391	NP_000225	dd_02463	1.9e-245 (0)
Hereditary nonpolyposis colorectal cancer, type 1 (HNPCC) (MSH2)	120435	AAA18643	dd_00995	2.4e-237 (1)
*Hyperinsulinism (ABCC8)	600509	Q09428	dd_00006	3.0e-220 (1)
G6PD deficiency (G6PD)	305900	NP_000393	dd_01534	5.1e-190 (0)
*Stargardt's (ABCA4)	601691	AAC51144	dd_02412	6.5e-189 (3)
Deafness, hereditary (MYO15)	602666	AAF05903	dd_02568	1.2e-182 (5)
Familial cardiac myopathy (MYH7)	160760	P12883	dd_02401	4.2e-177 (5)
Chediak–Higashi (CHS1)	214500	NP_000072	dd_02608	3.3e-151 (1)
Cancer (AKT2)	164731	AAA58364	dd_02928	1.5e-94 (2)
HNPCC (MSH3)	600887	AAB06045	dd_01030	8.9e-78 (1)

From a list of 287 confirmed human disease protein sequences[30] those are shown that match a *D. discoideum* chromosome 2 protein with a BLASTP probability of less than 1.0×10^{-50}, indicating a strong similarity. Only the best hit is listed and the total number of additional strong hits ($P < 1.0 \times 10^{-50}$) is given in parentheses after the probability score. OMIM, Online Mendelian Inheritance in Man (http://www.ncbi.nlm.nih.gov/omim/).
*Homologous proteins discussed in the text.

latter two genes and to several other human disease-related genes (Table 3).

What can the genome of *D. discoideum* tell us about the common genomic repertoire of eukaryotic life? To address this question, we compared the protein products of the 2,799 genes of chromosome 2 to the complete protein sets of fully sequenced eukaryotes ($P < 10^{-15}$) and found that 973 proteins (35%) have matches. Of these only 487 share similarities across plants (*A. thaliana*), fungi (*S. cerevisiae* and *S. pombe*) and metazoa (*C. elegans*, *D. melanogaster* and available vertebrate sequences). A surprisingly high number (141) have matches with metazoa but not plants or fungi (Fig. 3a). Subdividing metazoa into fly, worm and vertebrates shows that even amongst this closely related group not all genes have comparable similarities in each species (Fig. 3b). This may reflect gene losses during evolution, or evolutionary rate variations for identical genes in different organisms, but could also reflect a gain of function for specific gene groups in each organism. If chromosome 2 is taken as a representative quarter of the *D. discoideum* genome, then less than 2,400 different genes are shared between *D. discoideum*, *S. pombe*, *S. cerevisiae*, *C. elegans*, *A. thaliana*, *D. melanogaster* and man. This number might well represent the 'minimal gene set' of a free-living eukaryote. From random mutagenesis studies it was previously estimated that the essential genes of yeast comprise only about 30% of all its genes[25]. Our estimate of the number of genes shared by all eukaryotes is close to this number.

Our analyses are all predicated on the assumption that chromosome 2, representing 25% of the genome, is typical of the remainder. This seems a reasonable assumption and other evidence on the distribution of mapped genes[15] does not suggest that chromosome 2 is particularly atypical. Our findings can also clarify the evolutionary position of *D. discoideum*. Its genome exhibits greater similarities to metazoa than to plants or fungi (Fig. 3). This supports the finding of a recent phylogenetic analysis of conserved protein sequences which placed the Myxomycota (to which *D. discoideum* belongs) at a position before the branching of the metazoa and fungi but after the divergence of the plant kingdom[7]. *D. discoideum* does not appear to have suffered the extensive gene loss observed in *S. cerevisiae* and therefore its gene content may better represent a basic eukaryotic genome. This conservation of the complete gene set makes *D. discoideum* well suited for functional studies of genes not represented in yeast. Its surprisingly high gene number may in part reflect the higher order of complexity associated with multicellular life. □

Methods

Further information on sequence data and analysis results can be accessed via http://genome.imb-jena.de/dictyostelium/ and http://www.uni-koeln.de/dictyostelium/.

Sequencing and assembly

Library construction and sequencing was done as described previously[19]. 160,000 chromosome 2 library-derived reads were pooled with the nonrepetitive reads from other *D. discoideum* whole chromosome shotgun projects to give a total of 500,000 reads. The subset of reads matching genes mapped to chromosome 2 (ref. 15) were assembled to build seed contigs, and further contigs were assembled around complementary reads from the clones in these seeds (for details see http://genome.imb-jena.de/dictyostelium/chr2/seeds.html). The previously published[15] map order of the 'seed' genes was largely confirmed. Considering the combined data of the HAPPY map and sequence assembly, it is likely that the discrepancies arise from errors in the earlier YAC contigs, which have been shown to suffer from a proportion of misplaced clones[26]. The assembly database was then enlarged by the incorporation of reads with a higher than average frequency of occurrence in the chromosome 2 library reads (these are more likely to originate from chromosome 2 than from other chromosomes, owing to our preferential use of the chromosome 2 specific clone libraries). The contigs were extended by the incorporation of further reads which were found by BLAST analysis of the contig ends. This assembly method yielded about 1,100 contigs larger than 2 kb. Chromosomal assignment of each contig was checked on the basis of its content of sequences derived from each of the different chromosome-enriched libraries (K.S., unpublished software). In this way, contaminant sequences were filtered out; conversely, reads derived from the other whole chromosome shotgun projects but assigned to chromosome 2 were incorporated into our assembly. To ensure that we had not missed portions of the chromosome by this strategy we assembled all chromosome 2 library-derived reads and checked the resulting contigs for chromosome specificity. The resulting additional contigs were added to the chromosome 2-specific assembly. All contigs were manually inspected to ensure data accuracy. Clones spanning sequencing gaps between neighbouring contigs defined scaffolds. Directed closure of these gaps was done using custom primers to walk on existing clones. Additional gaps were closed by using the transposon insertion technique and polymerase chain reaction (PCR) approaches.

Mapping

As part of the ongoing genome-wide *D. discoideum* HAPPY mapping project, a short range (~100 kb), high-resolution (mean, 15 kb) HAPPY mapping panel was prepared from AX4 genomic DNA, pre-amplified by PEP (primer extension pre-amplification) and diluted before use as a template for marker typing. Hemi-nested primers were designed for 824 markers selected from the sequencing projects. Markers were typed onto the HAPPY mapping panel and sorted into linkage groups as previously described[26]. Maps for each linkage group were generated and validated by inspection to reduce the risk of incorporating spurious intermarker linkages. The results obtained so far define the order of 365 chromosome 2 markers in 12 large linkage groups (for details see http://genome.imb-jena.de/dictyostelium/chr2/linkage.html). Further positional information was obtained from PCR screening of a cYAC library with insert sizes of 80–100 kb covering the genome approximately sevenfold. Sequence contigs or HAPPY linkage groups were assumed to be linked if primer pairs derived from them hit the same cYAC clone(s). By integrating the HAPPY, cYAC and sequence scaffold data, a region spanning 6.5 Mb was robustly assembled. Only four sequence scaffolds, totalling 0.6 Mb, could not be placed onto the map. Of these, two are too large to fit in the gaps of the linked 6.5-Mb portion of the chromosome, and are presumably located at the ends. The assembly produced 71 unlinked orphan contigs amounting to 0.41 Mb, consisting mainly of fragments of complex repetitive elements.

Sequence analysis

G+C content was calculated using a sliding window of 10,000 bases and a step size of 1,000 bases. Strand-specific CDS density was measured as percentage of coding triplets in a stepped window of 5,000 bases. A database containing the complex repetitive elements of *D. discoideum* was used with RepeatMasker to scan the sequence for repeats[19]. tRNAs were detected using tRNAscan-SE[27]. To define the genes on chromosome 2, the gene prediction program GeneID[28] was trained with 140 known *D. discoideum* genes and its parameters adjusted to be able to define proper gene borders and intron positions. The lower limit for the gene length was 120 bases of coding sequence. EST matches were defined by BLAST with >98% identity, and word length of 32. The protein products of predicted genes were compared to the databases of completed genomes: The *Arabidopsis* Information Resource (http://www.arabidopsis.org/home.html), Wormpep (http://www.sanger.ac.uk/Projects/C_elegans/wormpep/), ftp://ftp.ebi.ac.uk/pub/databases/edgp/sequence_sets/, *Saccharomyces* Genome Database (http://genome-www.stanford.edu/Saccharomyces/), The *Schizosaccharomyces pombe* Genome Sequencing Project (http://www.sanger.ac.uk/Projects/S_pombe/), Ensembl Genome Browser (http://www.ensembl.org) as well as against SWISS-PROT entries and TrEMBL. They were also checked for the presence of InterPro domains using the InterPro database (http://www.ebi.ac.uk/interpro). Functional classification was done automatically using the GO classification system (http://www.geneontology.org/)[29].

Received 14 December 2001; accepted 26 April 2002; doi:10.1038/nature00847.

1. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274,** 546–567 (1996).
2. Wood, V. *et al.* The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415,** 871–880 (2002).
3. The *C. elegans* Sequencing Consortium Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282,** 2012–2018 (1998).
4. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–2195 (2000).
5. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
6. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
7. Baldauf, S. L., Roger, A. J., Wenk-Siefert, I. & Doolittle, W. F. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290,** 972–977 (2000).
8. Loomis, W. F. Genetic networks that regulate development in *Dictyostelium* cells. *Microbiol. Rev.* **60,** 135–150 (1996).
9. Eichinger, L., Lee, S. S. & Schleicher, M. *Dictyostelium* as model system for studies of the actin cytoskeleton by molecular genetics. *Microsc. Res. Technol.* **47,** 124–134 (1999).
10. Parent, C. A. & Devreotes, P. N. A cell's sense of direction. *Science* **284,** 765–770 (1999).
11. Firtel, R. A. & Meili, R. *Dictyostelium*: a model for regulated cell movement during morphogenesis. *Curr. Opin. Genet. Dev.* **10,** 421–427 (2000).
12. Noegel, A. A. & Schleicher, M. The actin cytoskeleton of *Dictyostelium*: a story told by mutants. *J. Cell Sci.* **113,** 759–766 (2000).
13. Kay, R. R. & Williams, J. G. The *Dictyostelium* genome project: an invitation to species hopping. *Trends Genet.* **15,** 294–297 (1999).
14. Cox, E. C., Vocke, C. D., Walter, S., Gregg, K. Y. & Bain, E. S. Electrophoretic karyotype for *Dictyostelium discoideum*. *Proc. Natl Acad. Sci. USA* **87,** 8247–8251 (1990).
15. Loomis, W. F. & Kuspa, A. Dictyostelium—*A Model System for Cell and Developmental Biology* (eds Maeda, Y., Inouye, K. & Takeuchi, I.) 15–30 (Universal Academic, Tokyo, 1997).
16. Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282,** 1126–1132 (1998).
17. Bowman, S. *et al.* The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400,** 532–538 (1999).
18. Glöckner, G. Large scale sequencing and analysis of AT rich eukaryote genomes. *Curr. Genom.* **1,** 289–299 (2000).

19. Glöckner, G. *et al.* The complex repeats of *Dictyostelium discoideum*. *Genome Res.* **11,** 585–594 (2001).
20. Dear, P. H. in *Genome Mapping—A Practical Approach* (ed. Dear, P. H.) 95–124 (IRL Press, Oxford, 1997).
21. Pan, W. C. & Blackburn, E. H. Single extrachromosomal ribosomal RNA gene copies are synthesized during amplification of the rDNA in *Tetrahymena*. *Cell* **23,** 459–466 (1981).
22. Lim, L. P. & Burge, C. B. A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl Acad. Sci. USA* **98,** 11193–11198 (2001).
23. Morio, T. *et al.* The *Dictyostelium* developmental cDNA project: generation and analysis of expressed sequence tags from the first-finger stage of development. *DNA Res.* **5,** 335–340 (1998).
24. Apweiler, R. *et al.* InterPro—an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16,** 1145–1150 (2000).
25. Goebl, M. G. & Petes, T. D. Most of the yeast genomic sequences are not essential for cell growth and division. *Cell* **46,** 983–992 (1986).
26. Konfortov, B. A., Cohen, H. M., Bankier, A. T. & Dear, P. H. A high-resolution HAPPY map of *Dictyostelium discoideum* chromosome 6. *Genome Res.* **10,** 1737–1742 (2000).
27. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).
28. Parra, G., Blanco, E. & Guigo, R. GeneID in *Drosophila*. *Genome Res.* **10,** 511–515 (2000).
29. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
30. Fortini, M. E., Skupski, M. P., Boguski, M. S. & Hariharan, I. K. A survey of human disease gene counterparts in the *Drosophila* genome. *J. Cell Biol.* **150,** F23–F30 (2000).

Supplementary Information accompanies the paper on *Nature*'s website (http://www.nature.com/nature).

Acknowledgements

We thank S. Förste, N. Zeisse, S. Rothe, S. Landmann, R. Schultz, S. Müller and R. Müller for expert technical assistance. We also thank the working team of the Japanese cDNA project (http://www.csm.biol.tsukuba.ac.jp/cDNAproject.html) for sharing data. The sequencing of chromosome 2 was supported by the Deutsche Forschungsgemeinschaft, with partial support by Köln Fortune. Additional support was obtained from the NIH, the Medical Research Council and the EU. Members of the Sanger Institute *Dictyostelium discoideum* sequencing team are listed at http://www.sanger.ac.uk/Projects/D_discoideum/team.shtml.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.A.N. (e-mail: noegel@uni-koeln.de) or G.G. (e-mail: gernot@imb-jena.de).

|| The *Dictyostelium* Genome Sequencing Consortium (members not included in the main author list):

Sequencing and Analysis:
The Sanger Institute Dictyostelium sequencing team (led by Bart G. Barrell & Marie-Adèle Rajandream)[1], Jeffrey G. Williams[2], Robert R. Kay[3], Adam Kuspa[4], Richard Gibbs[4], Richard Sucgang[4], Donna Muzny[4] & Brian Desany[4]

Generation of cYAC library:
Kathy Zeng[5], Baoli Zhu[5] & Pieter de Jong[5]

Advisory Committee for the DFG-funded project:
Theodor Dingermann[6], Günther Gerisch[7], Peter Philippsen[8], Michael Schleicher[9], Stephan C. Schuster[10] & Thomas Winckler[6]

1, The Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK; 2, University of Dundee, MSI/WTB Complex, Dundee, UK; 3, MRC Laboratory, of Molecular Biology, Cambridge CB2 2QH, UK; 4, Baylor College of Medicine, Houston, Texas 77030, USA; 5, Children's Hospital Oakland – BACPAC Resources, Oakland, California 94609, USA; 6, Institut für Pharmazeutische Biologie, Universität Frankfurt (Biozentrum), Frankfurt am Main, 60439, Germany; 7, Max-Planck-Institut für Biochemie, 82152 Martinsried, Germany; 8, Molecular Microbiology, Biozentrum der Universität, 4056 Basel, Switzerland; 9, A.-Butenandt-Institut/Zellbiologie, Ludwig-Maximilians-Universität, 80336 München, Germany; 10, Max-Planck-Institut für Entwicklungsbiologie, 72076 Tübingen, Germany

Intracellular calcium stores regulate activity-dependent neuropeptide release from dendrites

Mike Ludwig*, Nancy Sabatier*, Philip M. Bull*, Rainer Landgraf†, Govindan Dayanithi‡ & Gareth Leng*

** Department of Biomedical Sciences, University of Edinburgh Medical School, George Square, Edinburgh EH8 9XD, UK*
† Max Planck Institute of Psychiatry, Clinical Institute, Kraepelinstraße 2-10, 80804 Munich, Germany
‡ Department of Neurobiology, INSERM 432, University of Montpellier II, Place Eugene Bataillon, F-34094 Montpellier, Cedex 5, France

Information in neurons flows from synapses, through the dendrites and cell body (soma), and, finally, along the axon as spikes of electrical activity that will ultimately release neurotransmitters from the nerve terminals. However, the dendrites of many neurons also have a secretory role, transmitting information back to afferent nerve terminals[1–4]. In some central nervous system neurons, spikes that originate at the soma can travel along dendrites as well as axons, and may thus elicit secretion from both compartments[1]. Here, we show that in hypothalamic oxytocin neurons, agents that mobilize intracellular Ca^{2+} induce oxytocin release from dendrites without increasing the electrical activity of the cell body, and without inducing secretion from the nerve terminals. Conversely, electrical activity in the cell bodies can cause the secretion of oxytocin from nerve terminals with little or no release from the dendrites. Finally, mobilization of intracellular Ca^{2+} can also prime the releasable pool of oxytocin in the dendrites. This priming action makes dendritic oxytocin available for release in response to subsequent spike activity. Priming persists for a prolonged period, changing the nature of interactions between oxytocin neurons and their neighbours.

Neurons in the supraoptic nucleus (SON) of the hypothalamus project axons to the posterior pituitary, where oxytocin and vasopressin are secreted from axonal nerve terminals into the systemic circulation. These peptides are also released in large amounts from dendrites in the SON[5], but secretion at these two sites is not consistently correlated. Suckling evokes oxytocin release in the SON[6] before significant peripheral secretion, whereas after osmotic stimulation, SON oxytocin release lags behind peripheral secretion[7]. During lactation, in response to suckling, oxytocin cells discharge with brief, intense bursts[8]; these bursts release boluses of oxytocin into the circulation that result in milk let-down from the mammary glands. The bursting activity can be blocked by central administration of oxytocin antagonists[9], thus central as well as peripheral oxytocin is essential for milk let-down. It has been proposed that suckling evokes dendritic oxytocin release that acts in a positive feedback manner to evoke bursting[10].

Oxytocin mobilizes intracellular Ca^{2+} from thapsigargin-sensitive stores in oxytocin cells[11]. Here we tested the hypothesis that this might be critical for dendritic oxytocin release. In anaesthetized rats, we implanted a microdialysis probe into the SON to measure oxytocin release in response to systemic osmotic stimulation. In some of these experiments we applied thapsigargin directly to the SON through the dialysis probe. Thapsigargin caused a significant increase in SON oxytocin release that returned to control levels after washout. Subsequent systemic osmotic stimulation (2 ml of 1.5 M NaCl, intraperitoneal injection) caused a much larger release of oxytocin in thapsigargin-pretreated rats than in controls. Osmotically stimulated oxytocin secretion into the circulation was unaffected by exposure of one SON to thapsigargin (Fig. 1a, b).

To test whether thapsigargin potentiated spike-dependent release