INSTITUT PASTEUR

COURS D'ANALYSE DES GENOMES

ANNEE UNIVERSITAIRE 2005-2006

Mercredi 9 novembre

13:00-14:30 *Conférence* : **L'annotation *in silico* des séquences génomiques bactériennes** — **Claudine MEDIGUE** (Génoscope, Evry)

Mercredi 16 novembre

9:00-10:30 *Conférence* : **Transferts génétiques horizontaux chez les bactéries** — **Didier MAZEL** (Institut Pasteur)

10:45-12:15 *Conférence* : **Génomique comparative des levures : cycles sexuels et régions subtélomériques** — **Cécile FAIRHEAD** (Institut Pasteur)

Mercredi 23 novembre

9:00-10:30 *Conférence* : **Analyse des complexes protéiques par les approches protéomiques** — **Bertrand SERAPHIN** (CGM CNRS, Gif sur Yvette)

10:45-12:15 *Conférence* : **Approches protéomiques pour la mise au point de nouveaux médicaments anti-herpès** — **Jean-Jacques DIAZ** (INSERM U369, Faculté de Médecine Lyon)

Vendredi 25 novembre

9:00-10:30 *Conférence* : **Récents développements en génomique structurale** — **Arnaud DUCRUIX** (UMR 8015 CNRS, Paris)

10:45-12:15 *Conférence* : **Approche «transcriptome»: bases théoriques et applications** — **Catherine NGUYEN** (INSERM ERM206, Marseille)

Mardi 29 novembre

14:00-15:30 *Conférence* : **Le génome de la paramécie** — **Linda SPERLING** (CGM CNRS, Gif sur Yvette)

Mercredi 30 novembre

9:00-10:30 *Conférence* : **La génomique fonctionnelle chez la souris** — **Marie-Christine SIMMLER** (UMR 7622 CNRS, Paris)

10:45-12:15 *Conférence* : **Génomique et génétique humaine : les polymorphismes du génome humain comme outils pour l'étude des maladies** — **Laurence COLLEAUX** (INSERM U393 Necker, Paris)

Vendredi 2 décembre

9:00-10:30 *Conférence* : **Etudes de la réplication et de l'instabilité génomique sur molécules uniques d'ADN** — **Aaron BENSIMON** (Institut Pasteur)

10:45-12:15 *Conférence* : **L'évolution des génomes viraux** — **Simon WAIN-HOBSON** (Institut Pasteur)

Lundi 5 décembre

9:00-10:30 *Conférence* : **Phylogénomique des *Archaea*** — **Patrick FORTERRE** (Institut Pasteur)

10:45-12:15 *Conférence* : **Reconstruction du génome d'un vertébré ancestral** — **Hugues ROEST-CROLLIUS** (CNRS UMR 8541, ENS, Paris)

Mardi 6 décembre

9:00-10:30 *Conférence* : **Biogenèse des mitochondries : une approche génomique** — **Claude JACQ** (CNRS UMR 8541, ENS, Paris)

Vendredi 9 décembre

9:00-10:30 *Conférence* : **Comment la génomique peut aider à comprendre le pouvoir pathogène de la bactérie *Helicobacter pylori* ?** — **Hilde DE REUSE** (Institut Pasteur)

10:45-12:15 *Conférence* : **Du génome à la cellule : comment prédire la fonction des gènes ?** — **Antoine DANCHIN** (Institut Pasteur)

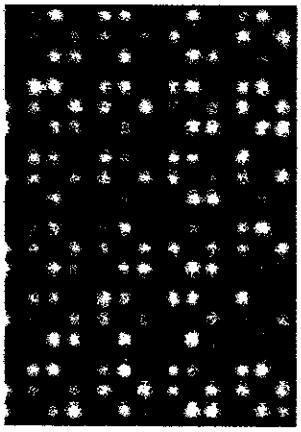
N.B : Les conférences du Cours d'Analyse des Génomes ont lieu salle «R. LEGROUX» du Centre d'Enseignement de l'Institut Pasteur (bâtiment A. YERSIN

MEDECINE/SCIENCES 2005; 21: 539-44

Les puces à ADN vont-elles révolutionner l'identification des bactéries ?

Philippe Glaser

Les puces à ADN sont des multicapteurs permettant de caractériser et quantifier un acide nucléique dans un échantillon. Elles apportent une solution innovante au problème ancien de la détection, de l'identification et du typage de bactéries dans un échantillon. Elles permettent la caractérisation génomique rapide de bactéries pathogènes et facilitent les études épidémiologiques, par exemple pour le contrôle des maladies nosocomiales ou la surveillance du bioterrorisme. Ces puces sont développées dans les laboratoires de recherche pour l'étude de la diversité et de l'évolution du monde bactérien, pour rechercher des gènes de résistance aux antibiotiques et pour la caractérisation de communautés bactériennes constituées de centaines d'espèces. L'industrialisation du processus de fabrication et d'utilisation, rendant la technologie robuste tout en en diminuant son coût, devrait permettre son utilisation dans les laboratoires hospitaliers et d'analyses spécialisées, puis sa généralisation aux laboratoires de ville. <



Unité de génomique des micro-organismes pathogènes, Institut Pasteur, 28, rue du Docteur Roux, 75724 Paris Cedex 15, France.
pglaser@pasteur.fr

Depuis la démonstration par Koch, en 1876, que la maladie du charbon est due à *Bacillus anthracis* [1], l'identification des microbes responsables des maladies infectieuses a été une priorité de la microbiologie clinique. L'isolement du germe et sa description phénotypique sont nécessaires pour mieux comprendre la maladie qu'il provoque et pour un diagnostic sûr. L'épidémiologie des maladies infectieuses cherche à élucider les mécanismes de transmission des agents infectieux, l'existence de réservoirs et leur origine. L'identification de l'espèce n'est alors pas suffisante et la caractérisation plus fine (le typage) de la bactérie isolée est nécessaire afin de déterminer la probabilité pour deux isolats d'avoir la même origine. Répondre à cette question a des implications multiples: pour les maladies nosocomiales, ces recherches doivent permettre d'identifier l'origine hospitalière de l'infection et de mettre en place les procédures pour y remédier [2]; pour des maladies d'origine alimentaire ou environnementale, comme la listériose ou la légionellose, des réseaux de surveillance[1] cherchent à identifier par des études épidémiologiques la source de la contamination, afin de l'éliminer. La démonstration de l'origine d'une infection a et aura de plus en plus des implications légales, et le typage des bactéries incriminées est un des éléments de l'action judiciaire. La crainte du bioterrorisme pousse également à rechercher des outils performants pour identifier les germes dispersés dans l'environnement, pour retrouver l'origine des souches et comprendre les éventuels trafics de ces armes biologiques.

Des méthodes de microbiologie classique développées au XIX[e] siècle comme la coloration de Gram aux analyses moléculaires récentes fondées sur l'analyse des acides nucléiques, les microbiologistes ont développé de très nombreuses méthodes d'identification et de typage des

Article reçu le 13 décembre 2004 et accepté le 7 mars 2005.

[1] Voir par exemple l'organisation EWGLI pour la légionellose (http://www.ewgli.org/), le réseau européen med-vet-net pour la prévention des zoonoses (http://www.med-vetnet.org) ou, aux États-Unis, le réseau Pulsenet pour les infections d'origine alimentaire (http://www.cdc.gov/pulsenet).

bactéries, mettant à profit leur diversité [3]. L'objectif de cet article est de montrer le formidable potentiel des puces à ADN dans ce domaine, mais aussi d'analyser les raisons pour lesquelles cette technologie tarde à s'implanter dans les laboratoires d'analyse microbiologique.

La diversité du monde bactérien

La caractérisation de la diversité bactérienne sous-tend toutes les méthodes d'identification. Au cours de la dernière décennie, notre conception de cette diversité a été transformée par l'analyse des génomes. Au mois d'août 1995, la communauté scientifique a été surprise par la publication de la séquence du génome d'*Haemophilus influenzae* [4]. La séquence du génome d'une bactérie modèle était attendue, mais c'est une bactérie d'importance clinique, peu étudiée, pour laquelle n'étaient disponibles ni carte physique ni carte génétique, qui a vu la première l'ensemble de ses gènes décryptés. La démonstration que le séquençage d'un génome bactérien pouvait être réalisé rapidement a ouvert la porte à la systématisation de son application aux bactéries pathogènes, qui représentent la majorité des plus de 180 séquences génomiques publiées à ce jour[2]. La connaissance du génome permet le développement de nouvelles méthodes de typage moléculaire, mais aussi phénotypique, par la découverte d'activités enzymatiques spécifiques.

Cependant, le décryptage du génome d'un isolat n'est pas suffisant pour connaître une espèce. La disponibilité de séquences génomiques de plusieurs souches a permis de mettre en évidence la diversité génomique d'une espèce, non seulement au niveau du polymorphisme de chaque gène, mais aussi du répertoire de gènes présents. Il est possible de distinguer dans un génome un squelette conservé entre tous les isolats d'une espèce et un ensemble d'îlots qui sont spécifiques d'un clone ou d'un lignage particulier [5]. Cette partie variable du génome est constituée d'éléments mobiles comme des bactériophages, des plasmides ou des transposons, mais également de groupes de gènes insérés ou délétés indépendamment de tout système de recombinaison spécifique. Ces gènes apportent des fonctions particulières à un clone, comme des fonctions métaboliques, de virulence, de production de toxines ou de résistance à des antibiotiques. Cette diversité génomique explique aussi la diversité de virulence retrouvée au sein d'une espèce. Les méthodes de typage moléculaire sont fondées sur le polymorphisme des séquences et la diversité du contenu génétique des génomes; les puces à ADN permettent d'analyser ces deux aspects.

Méthodes d'identification et de typage

L'identification bactérienne nécessite l'isolement de la bactérie sous forme d'une colonie. Cette première étape peut être difficile si l'échantillon biologique n'est pas normalement stérile, comme un prélèvement de gorge ou un prélèvement de selles. Traditionnellement, l'identification de l'espèce se faisait en combinant l'observation microscopique et l'analyse phénotypique, en étudiant la forme et la couleur des colonies et les propriétés métaboliques et enzymatiques. Cette méthode peut être remplacée par une analyse moléculaire soit en utilisant des tests immunologiques, soit sur la base de séquences d'ADN, par exemple après amplification par PCR. Les méthodes de PCR présentent l'avantage de pouvoir être réalisées directement à partir d'un échantillon biologique. Le gène codant pour l'ARN ribosomique 16S est un des marqueurs d'espèce les plus utilisés : il est en effet possible de définir des amorces universelles pour son amplification, et la comparaison de sa séquence avec des banques de données de référence permet de déterminer l'espèce [6].

De la même manière, après l'identification de l'espèce, les premières méthodes de typage bactérien étaient phénotypiques, par comparaison des antigènes de surface (sérotypie) ou analyse de la sensibilité à des bactériophages (lysotypie). Durant ces vingt dernières années, de nombreuses méthodes de typage moléculaire ont été développées [7]. La valeur de ces méthodes dépend de leur pouvoir de discrimination entre les isolats, de leurs facilité et rapidité d'utilisation et de la reproductibilité des résultats entre différents laboratoires. Cette reproductibilité est essentielle pour l'échange des données, leur organisation sous forme de bases de données internationales et la mise en place de systèmes de surveillance épidémiologique performants. Les méthodes considérées aujourd'hui comme les plus fiables sont l'analyse des fragments de restriction de l'ADN chromosomique après migration en champ pulsé (pulsotype) et l'analyse des séquences nucléotidiques de plusieurs gènes de ménage (exprimés dans toutes les cellules) (STML, séquençotypage multilocus, MLST en anglais). Le pulsotype dépend du polymorphisme des sites de reconnaissance par les enzymes de restriction et de l'organisation du génome. Le type MLST dépend uniquement de la vitesse d'évolution nucléotidique au sein de l'espèce et d'éventuels transferts génétiques horizontaux.

[2] http://www.genomesonline.org/

Puces à ADN

Le principe d'une puce à ADN réside dans la reconnaissance, c'est-à-dire l'hybridation, entre deux molécules d'ADN simple brin complémentaires *(Figure 1)*. L'échantillon (ADN ou ARN), marqué de manière fluorescente, est mis en contact avec la puce portant plusieurs milliers de sondes qui sont des fragments d'ADN ou des oligonucléotides de séquence connue [8, 9]. Après lavage du matériel fixé de manière non spécifique, le signal est quantifié au niveau de chaque sonde. Sa valeur dépendra de la concentration en molécules marquées complémentaires de la sonde dans l'échantillon et du degré de complémentarité (le pourcentage d'identité) avec la sonde.

Actuellement, l'utilisation la plus courante des puces à ADN concerne la quantification des ARN messagers d'un échantillon (transcriptome) afin de comparer les profils de transcription de deux échantillons obtenus dans des conditions de croissance différentes. Dans le cadre de l'identification bactérienne et du typage, leur utilisation est différente: ces puces permettent la détection d'une séquence d'ADN dans un mélange, d'identifier un polymorphisme de séquence (SNP) et de re-séquencer un fragment d'ADN *(Figure 2)*.

Les puces à ADN comme outils de détection

Des puces à ADN sont utilisées pour détecter un produit de PCR et remplacent ainsi la migration sur gel en validant la spécificité de l'amplification par sa complémentarité avec la sonde. Pour l'analyse d'un échantillon simple, cette procédure est peu compétitive par rapport à la PCR en temps réel ou à l'électrophorèse capillaire. En revanche, dans le cas d'un échantillon complexe, les puces à ADN permettent l'analyse simultanée d'un grand nombre de fragments. Des puces permettant d'identifier les bactéries présentes dans un échantillon sur la base de l'analyse des séquences d'ARN ribosomique 16S sont en développement dans plusieurs institutions *(Figure 2A)*. L'ADN total est extrait à partir d'un échantillon et l'ensemble des ADN codant pour les ARN 16S sont amplifiés, en utilisant des amorces universelles, et hybridés sur des puces portant des sondes spécifiques de l'ARN 16S des espèces recherchées. Cet outil est développé pour la surveillance du bioterrorisme avec un ensemble de sondes spécifiques des agents pathogènes potentiellement utilisés.

Les puces à ADN pour le re-séquençage

La connaissance de la séquence complète d'un génome est le niveau ultime de typage, mais, dans le contexte actuel, le séquençage de chaque isolat n'est pas envisageable. Des puces oligonucléotides sont donc développées pour obtenir des informations partielles sur les génomes bactériens *(Figure 2B)*. Des puces sont utilisées pour le typage par MLST de *Staphylococcus aureus* [10]. Les sept locus analysés sont amplifiés par PCR et hybridés à la puce au lieu d'être séquencés un par un. La société BioMérieux, en collaboration avec Affymetrix, a été pionnière dans le domaine en développant des puces pour l'identification de *Mycobacterium tuberculosis* et la recherche de mutations entraînant la résistance à l'isoniazide et à la rifampicine [11]. L'amélioration de la sensibilité de la technique doit

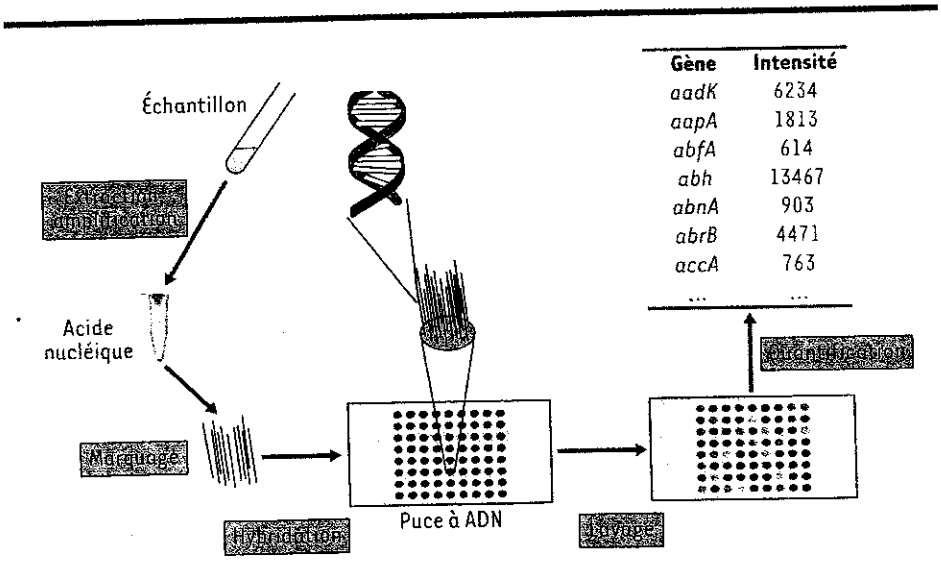


Figure 1. Analyse d'acide nucléique par puce à ADN. L'ADN ou l'ARN est purifié et éventuellement amplifié à partir d'un échantillon biologique. Il est ensuite marqué de manière fluorescente et mis en contact avec les sondes portées par la puce. Lors de cette étape d'hybridation, les acides nucléiques marqués vont s'apparier avec les sondes ADN fixées sur le support. Une étape de lavage permet ensuite d'éliminer les acides nucléiques marqués fixés de manière non spécifique. Finalement, la fluorescence au niveau de chaque dépôt de sonde sera quantifiée au moyen de tubes photomultiplicateurs ou d'une caméra CCD *(charge-coupled device)*. Les valeurs obtenues pour chaque sonde, comme indiquées sur le tableau, doivent ensuite être traitées aux moyens d'outils informatiques pour obtenir la caractérisation de l'échantillon.

permettre d'utiliser l'ADN génomique total pour détecter un ensemble de positions polymorphes réparties sur le génome. Il serait alors possible d'avoir une vision globale du génome et de pointer sur des mutations particulières, comme celles entraînant la résistance à un antibiotique.

Les puces à ADN pour la caractérisation génomique

Les puces à ADN sont aussi utilisées pour caractériser la partie variable du génome d'un clone. Ces puces « biodiversité » portent des sondes correspondant à des gènes qui ne sont pas présents dans tous les isolats d'une espèce *(Figure 2C)*. Elles sont établies à partir de la comparaison des séquences de plusieurs génomes. Par une seule expérience d'hybridation, ces puces permettent d'établir une véritable empreinte digitale correspondant aux gènes présents ou absents dans un clone. Ainsi, à la différence de la majorité des méthodes de typage, les puces à ADN apportent une information fonctionnelle sur la nature des gènes qui différencient deux isolats. Ces résultats peuvent être comparés aux données phénotypiques sur les souches, notamment en relation avec leur virulence. De telles puces de typage ont été établies pour *Listeria monocytogenes*, sur la base de la séquence génomique de deux isolats et d'un isolat de *L. innocua*, une espèce non pathogène très proche de *L. monocytogenes* [12], et pour *S. aureus*, en combinant les connaissances de sept génomes et en ajoutant des gènes de résistance aux antibiotiques ainsi que des gènes codant pour des toxines [13]. Dans les deux cas, l'analyse d'une collection de souches par hybridation de ces puces à ADN a montré qu'elles étaient un outil de typage puissant, aussi résolutif que l'analyse en champ pulsé. Par ailleurs, pour *S. aureus*, ces puces se sont révélées extrêmement efficaces pour l'identification des résistances aux antibiotiques, avec un très bon accord entre les résultats d'hybridation et les résultats d'antibiogramme. Finalement, la confrontation des données de phylogénie, fondée sur une analyse par MLST et la distribution des gènes entre les isolats, permet de mettre en évidence des phénomènes de transferts génétiques horizontaux qui peuvent jouer un rôle important dans l'émergence de clones hypervirulents.

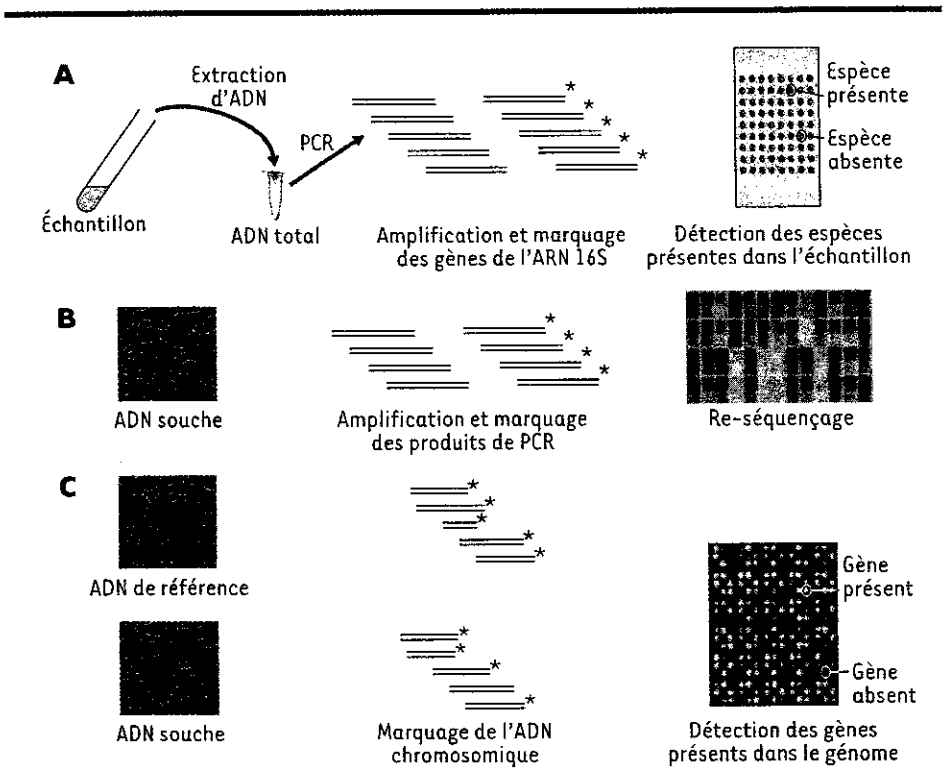


Figure 2. Trois applications des puces à ADN dans l'analyse microbiologique. A. Analyse d'une population bactérienne mixte. Après extraction de l'ADN, les régions codant pour l'ARN ribosomique 16S sont amplifiées pour toutes les bactéries de l'échantillon en utilisant des amorces universelles correspondant à des régions conservées dans toutes les espèces. La puce porte des oligonucléotides de séquence spécifique de chaque espèce. Après hybridation, la quantification du signal permet de détecter la présence de bactéries des différentes espèces représentées sur la puce et d'évaluer leurs quantités relatives. ***B.*** Puce Affymetrix de re-séquençage (http://www.affymetrix.com/). Ces puces portent plusieurs centaines de milliers d'oligonucléotides synthétisés *in situ* par une technique de photolithographie. Pour chaque base, quatre oligonucléotides sont synthétisés d'après une séquence de référence avec à la position centrale l'une des quatre bases, A, C, G ou T. La comparaison des signaux de fluorescence pour ces quatre oligonucléotides permet de déterminer la séquence à cette position. ***C.*** Détection de régions chromosomiques spécifiques d'un isolat. L'ADN de la souche à analyser et un ADN de référence sont marqués par des fluorophores ayant des propriétés spectrales différentes. La puce à ADN est hybridée avec un mélange des deux ADN marqués. Après analyse aux deux longueurs d'onde d'émission de fluorescence, les sondes absentes de la souche analysée n'émettront que pour la longueur d'onde du fluorophore de l'ADN de référence.

Transfert vers les laboratoires d'analyse

Les résultats obtenus dans les laboratoires de recherche montrent l'apport des puces à ADN en épidémiologie moléculaire et leur potentiel pour de nombreuses recherches en microbiologie. Théoriquement, c'est une technologie qui peut facilement être automatisée et industrialisée. Pourtant, alors que l'utilisation des puces Affymetrix pour *M. tuberculosis* a été publiée en 1997, cette technologie reste très confidentielle et il n'existe pas de produit de typage, fondé sur les puces à ADN, commercialisé à grande échelle. La première raison à ce délai est technologique. Les méthodes développées sont encore contraignantes et l'obtention de résultats reproductibles nécessite un ADN génomique d'une grande qualité. Des améliorations techniques accompagnées d'une simplification de la méthode augmentant sa robustesse sont nécessaires. La seconde raison à ce délai est liée à la difficulté de modifier des procédures reconnues au niveau international. Il est nécessaire qu'une nouvelle méthode soit d'abord validée et utilisée par les centres nationaux de référence et démontre sa supériorité avant de voir son utilisation systématisée dans les laboratoires d'analyse. La mise en place de sites Web équivalents à ceux qui ont été développés pour le MLST[3] devrait aussi promouvoir cette méthodologie. Les approches génomiques d'identification présentent un potentiel d'applications multiples et il est très probable que les problèmes technologiques seront résolus avec l'industrialisation du processus, en s'accompagnant d'une baisse des coûts. L'identification bactérienne devrait profiter des développements liés à l'utilisation des puces dans d'autres domaines de la santé, comme l'analyse des tumeurs ou de la prédisposition à certaines maladies. En retour, les développements réalisés en microbiologie clinique auront des répercussions en microbiologie de l'environnement et en microbiologie alimentaire.

Conclusions et perspectives

Les progrès de la génomique et la prise en compte de la génétique des populations pour les bactéries pathogènes permettent de mieux comprendre l'évolution et la diversité au sein des espèces pathogènes. La combinaison d'un outil performant de caractérisation des isolats bactériens et de bases de données internationales ouvertes incluant des isolats d'origines très diverses est l'occasion d'établir de nouveaux liens entre la surveillance microbiologique, la recherche clinique et la recherche fondamentale. Outre la caractérisation des génomes bactériens, l'application des puces à ADN pour l'analyse du transcriptome a contribué de manière très significative aux progrès récents dans l'étude du processus infectieux par une meilleure compréhension de la réponse de l'hôte et du parasite et des communications qui s'instaurent entre les deux partenaires au cours de la maladie [14]. ◊

SUMMARY

DNA-arrays, a breakthrough in bacterial identification?

DNA-arrays are mainly known for their application in transcriptome analysis leading for instance to the discovery of new marker genes for diagnostics and prognostics in oncology. However, DNA arrays are also used for massively parallel analysis of DNA molecules allowing their quantification, the detection of single nucleotide polymorphisms and re-sequencing. This multi detection system is now applied to the «old» problems of detecting and identifying bacteria in a biological sample and for the fine molecular characterization of a bacterial isolate. This new tool should serve for the diagnostic of an infection and for epidemiological studies such as those performed for the control of nosocomial infections or for the surveillance of bioterrorism attacks. DNA arrays carrying probes for 16S RNA specific of hundreds of bacterial species allow the identification of bacteria within a community by a single hybridization of amplified 16S rDNAs with universal primers and re-sequencing DNA arrays are used for multi locus sequence typing in a single step. Finally, the genome of an isolate could be characterized by DNA-arrays focused on a specific question like presence of toxin or antibiotic resistance genes. Up to now, DNA arrays are used in research laboratories for the rapid characterization at the genomic level of a strain collection, for evolutionary and population genetics studies and for the characterization of bacterial communities. Industrializing the process of DNA-array construction and hybridization is now needed in order to transfer this technology to hospitals and diagnostic laboratories. ◊

[3] http://www.mlst.net/

RÉFÉRENCES

1. Koch R. Die Aetiologie der Milzbrand-Krankheit, begründet auf die Entwicklungsgeschichte des *Bacillus anthracis*. *Beiträge zur Biologie der Pflanzen* 1876; 1: 277-308.
2. Struelens M. Molecular typing: a key tool for the surveillance and control of nosocomial infection. *Curr Opin Infect Dis* 2002; 15: 383-5.
3. Raoult D, Fournier PE, Drancourt M. What does the future hold for clinical microbiology? *Nat Rev Microbiol* 2004; 2: 151-9.
4. Fleischmann RD, Adams MD, White O, *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995; 269: 496-512.
5. Lan R, Reeves PR. Intraspecies variation in bacterial genomes: the need for a species genome concept. *Trends Microbiol* 2000; 8: 396-401.
6. Stackebrandt E, Goebel BM. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Bacteriol* 1994; 44: 846-9.
7. Van Belkum A. High-throughput epidemiologic typing in clinical microbiology. *Clin Microbiol Infect* 2003; 9: 86-100.
8. Schena M, Shalon D, Heller R, *et al.* Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci USA* 1996; 93: 10614-9.
9. Southern EM. DNA chips: analysing sequence by hybridization to oligonucleotides on a large scale. *Trends Genet* 1996; 12: 110-5.
10. Van Leeuwen WB, Jay C, Snijders S, *et al.* Multilocus sequence typing of *Staphylococcus aureus* with DNA array technology. *J Clin Microbiol* 2003; 41: 3323-6.
11. Troesch A, Nguyen H, Miyada CG, *et al.* *Mycobacterium* species identification and rifampin resistance testing with high-density DNA probe arrays. *J Clin Microbiol* 1999; 37: 49-55.
12. Doumith M, Cazalet C, Simoes N, *et al.* New aspects regarding evolution and virulence of *Listeria monocytogenes* revealed by comparative genomics. *Infect Immun* 2004; 72: 1072-83.
13. Trad S, Allignet J, Frangeul L, *et al.* DNA macroarray for identification and typing of *Staphylococcus aureus* isolates. *J Clin Microbiol* 2004; 42: 2054-64.
14. Bryant PA, Venter D, Robins-Browne R, Curtis N. Chips with everything: DNA microarrays in infectious diseases. *Lancet Infect Dis* 2004; 4: 100-11.

TIRÉS À PART

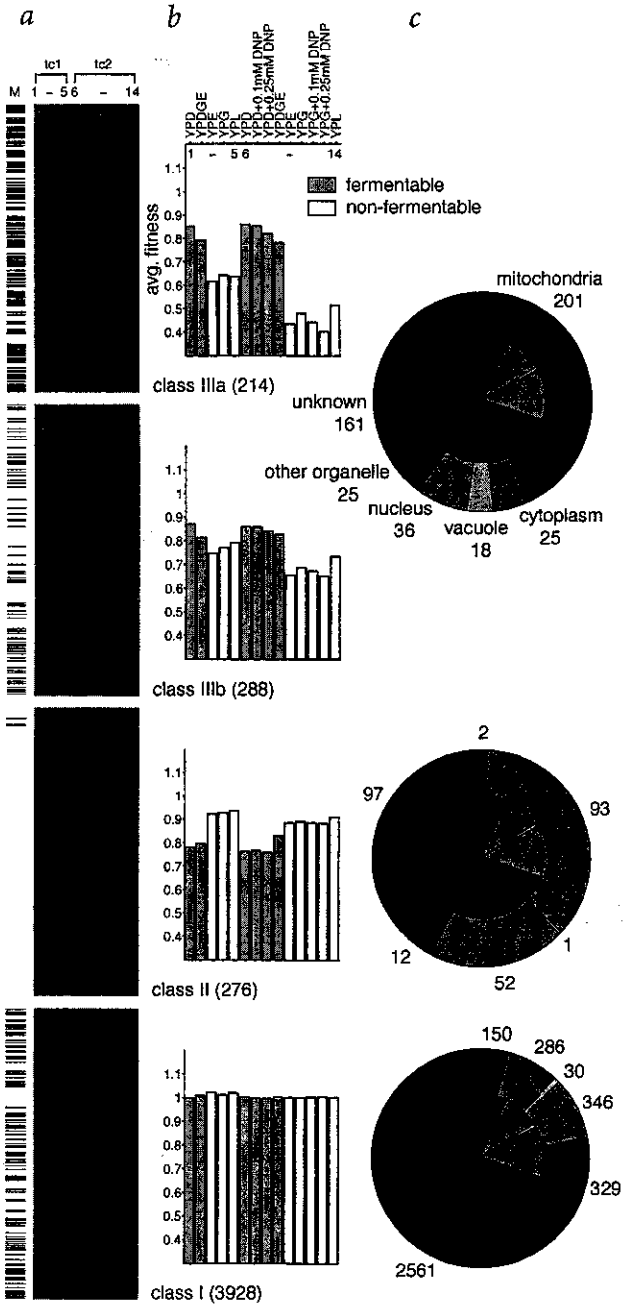
P. Glaser

Systematic screen for human disease genes in yeast

Lars M. Steinmetz[1,3]*, Curt Scharfe[2,3]*, Adam M. Deutschbauer[1], Dejana Mokranjac[4], Zelek S. Herman[3], Ted Jones[3], Angela M. Chu[2], Guri Giaever[3], Holger Prokisch[4], Peter J. Oefner[2,3] & Ronald W. Davis[1–3]

*These authors contributed equally to this work.

Published online: 22 July 2002, doi:10.1038/ng929



High similarity between yeast and human mitochondria allows functional genomic study of *Saccharomyces cerevisiae* to be used to identify human genes involved in disease[1]. So far, 102 heritable disorders have been attributed to defects in a quarter of the known nuclear-encoded mitochondrial proteins in humans[2]. Many mitochondrial diseases remain unexplained, however, in part because only 40–60% of the presumed 700–1,000 proteins involved in mitochondrial function and biogenesis have been identified[3]. Here we apply a systematic functional screen using the pre-existing whole-genome pool of yeast deletion mutants[4–6] to identify mitochondrial proteins. Three million measurements of strain fitness identified 466 genes whose deletions impaired mitochondrial respiration, of which 265 were new. Our approach gave higher selection than other systematic approaches, including fivefold greater selection than gene expression analysis. To apply these advantages to human disorders involving mitochondria, human orthologs were identified and linked to heritable diseases using genomic map positions.

The yeast deletion collection is a quantitative tool for systematically measuring the contribution to survival and reproduction (fitness) of most genes in the yeast genome[7]. In this study, 5,791 heterozygous diploid and 4,706 homozygous diploid deletion strains were quantitatively measured and monitored in parallel in 9 different medium conditions. Adapting classic diagnostic tests of mitochondrial function, we measured the growth (fitness) of mutant strains on non-fermentable substrates (including glycerol, lactate and ethanol) and compared

Fig. 1 Categorization of the whole genome according to phenotypes associated with gene deletions. *a*, Clustergram showing four fitness patterns found among the 4,706 homozygous diploid deletion strains in the pool. Each row represents a strain with a deletion of a different gene and each column a different medium condition. The number of strains in each cluster is indicated in parentheses next to the cluster name. For each strain, fitness values are indicated with a color scale ranging from blue to red, with blue representing levels below, and red levels above, the strain's median. Conditions 1–5 represent measurements from the first experimental time course (tc1) and columns 6–14 those from a repeat experiment (tc2). Marked in black at left are 353 strains with deletions of genes previously known to localize to or function in mitochondria (M). *b*, The bar graphs show the average fitness profiles for each cluster. For each condition, the height of the bar represents the growth rate of strains in the cluster relative to the average growth rate of the pool. Values of 1.0 indicate no difference, those less than 1.0 strains that grow more slowly than, and those greater than 1.0 strains that grow more quickly than, the pool average. The medium conditions, indicated above the graphs, are in the same order as the columns in the clustergram. *c*, The outer pie chart shows the composition of genes represented in each cluster according to MIPS localization categories[22], after the removal of all spurious ORFs. The inner pie charts represent the distribution over the genome. Because of the similarity in pattern, the class III clusters were combined.

[1]*Department of Genetics and* [2]*Department of Biochemistry, Stanford University School of Medicine, Stanford, California 94305, USA.* [3]*Stanford Genome Technology Center, Palo Alto, California 94304, USA.* [4]*Institute of Physiological Chemistry, University of Munich, Munich, Germany. Correspondence should be addressed to C.S. (e-mail: curts@stanford.edu) and L.M.S. (e-mail: larsms@stanford.edu).*

it with growth on fermentable sugar (glucose). Mutants with respiratory defects have impaired growth on non-fermentable substrates and are classically defined as 'petite'[8,9].

Of the 425 previously known genes encoding proteins involved in mitochondrial function and biogenesis[10], we detected the deletion strains for 353 in the homozygous diploid deletion pool (Fig. 1*a*). Of the remainder, the deletions were either lethal (37) or not successfully made (9), or the strains yielded signals too low to be considered detectable in the pool (26). Fifty-seven percent (201 of 353) of the mutants showed defects in growth on non-fermentable substrates, suggesting that about half of all mitochondrial-related proteins are essential for optimal respiratory activity[11] and can therefore be identified by a quantitative growth selection screen. The defects affected functions including oxidative phosphorylation, the tricarboxylic acid (TCA) cycle, mitochondrial protein synthesis and transport, ionic homeostasis and the metabolism of vitamins, cofactors and prosthetic groups. Mitochondrial genes whose deletion did not result in growth defects (152) encoded proteins with functions redundant or secondary to respiration: outer membrane transport, nucleotide transport and amino-acid metabolism.

In the heterozygous diploid pool, we observed few growth deficiencies, suggesting that mitochondrial proteins and enzymes were expressed in excess of minimum levels. In a small number of cases, including four strains heterozygous for deletions of known mitochondrial proteins (Rim2, Atp16, Nam9 and Mrpl19), we observed defects in growth on non-fermentable substrates, and the corresponding homozygous deletion mutant was lethal or not detected.

To identify new mutants with deficiencies in growth on non-fermentable substrates, we clustered the 4,706 quantitative homozygous diploid fitness profiles into four classes. Class I comprised 3,928 mutants (83.5%) with equal fitness on all the nutrients tested; class II comprised 276 with a higher fitness on non-fermentable substrates; and classes IIIa and IIIb comprised 502 with more and less severe defects, respectively, in fitness on non-fermentable substrates (Fig. 1*a*,*b*). We focused on class III for further analysis.

As physically overlapping genes cannot be distinguished by deletion analysis, we eliminated overlapping open reading frames (ORFs) that were defined as spurious[12], thereby reducing the number of genes in class III to 466. Because in these cases we detected the same phenotype for both overlapping ORFs, the spurious ORFs served as internal controls and validated our screen. Of the 466 genes, 201 (43%) encoded proteins with known mitochondrial localization or function[10]; 104 (22%), proteins that localized outside the mitochondria with functions in vacuolar and ion transport, transcription, and protein targeting, sorting and translocation; and 161 (35%), proteins with unknown subcellular localization, and in most cases unknown function (Fig. 1*c*).

We suspected that about half of the 161 proteins designated as unknown were localized to the mitochondria. Fifty-one had a putative mitochondrial import sequence[13], and 20 were homologous to sequences in *Rickettsia prowazekii*, which may be the closest ancestor of mitochondria[14] and has a total of 110 homologs of class III genes. Five of the 161 physically interacted with a known mitochondrial protein[15], and 21 had been identified as mitochondrially localized in a recent high-throughput immunolocalization study[16]. For six unknown proteins that each had a putative mitochondrial import sequence and were encoded by a gene whose deletion produced a severe phenotype, we assessed mitochondrial import directly using radiolabeled precursor proteins and isolated yeast mitochondria. Five of the six (Rsm18, Ygr101w, Yil157c, Mhr1 and Ppt2) were post-translationally imported into the mitochondria in a membrane potential–dependent manner. For Rsm18, Ygr101w and Yil157c, import was accompanied by removal of the signal peptide (Fig. 2).

Most mitochondrial proteins are encoded by the nuclear genome and are dependent on cellular protein expression and transport. The additional finding of nuclear (36) and cytoplasmic proteins (25) with defects in growth on non-fermentable substrates was therefore expected[2,17] (Fig. 1*c*). These genes integrate mitochondria into the cellular network, and their discovery illustrates an advantage of functionally based screening. Finding 68 cytoplasmic ribosomal proteins and 52 nuclear proteins whose absence led to deficient growth on fermentable substrates (class II) confirmed that the demand for translation[18] and transcription[19] is higher during fermentation than during respiration. In addition, combinations of defects in growth on fermentable and non-fermentable substrates were observed for genes of the glycolytic metabolic pathway (Fig. 3).

Two lines of evidence further confirmed the specificity of the quantitative deletion screen. First, for most proteins of the TCA cycle and respiratory chain, the corresponding deletion strains had defects in growth on non-fermentable substrates (Fig. 3). Second, genes encoding known mitochondrial proteins were 6.1-fold more enriched in our candidate set (class III) than in the genome (Fig. 1*c*). In comparison, gene-expression analysis of the diauxic shift[19] revealed an enrichment factor of only 1.2. As the diauxic shift represents a change in carbon source from glucose to ethanol and hence a shift from fermentation to respiration, the conditions are comparable to those of our growth selection screen. The enrichment data suggested that deletion phenotype is a more specific measure of gene function than is expression level. Twenty-four percent of genes whose deletion resulted in a

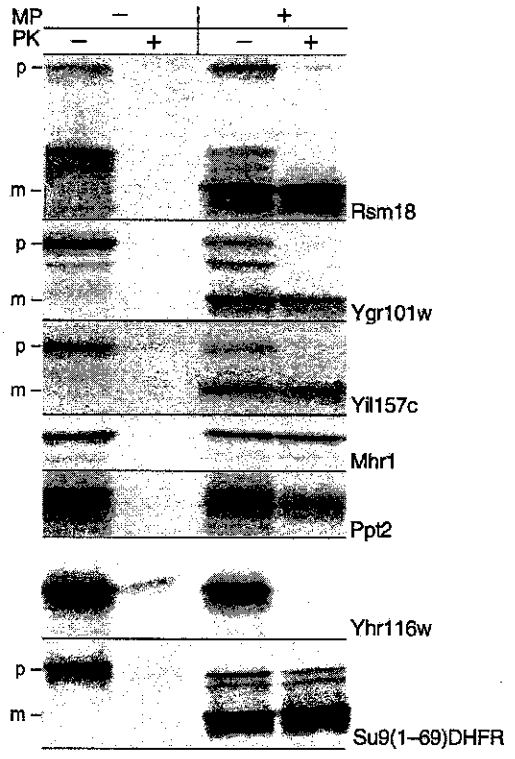


Fig. 2 Verification of mitochondrial candidates by import. Samples were derived by incubating radiolabeled proteins with mitochondria in the presence or absence of a membrane potential (MP) and the presence or absence of proteinase K (PK). In cases where import was accompanied by removal of the signal peptide, the precursor protein is labeled p; m indicates mature protein; Su9(1–69)DHFR, positive control.

defect in growth on non-fermentable substrates displayed at least a twofold difference in expression; conversely, 7% of genes with changing expression showed such a growth defect.

Rather than being used to infer the function of a gene from its expression level, expression analysis can also provide a detailed molecular phenotype, or signature profile. This approach produced a better agreement with our quantitative deletion screen. When we considered expression as a phenotype, 14 of 24 deletions of known mitochondrial proteins were clustered into the same group because they had a statistically similar signature profile[20]; in turn, our approach identified the same 14, as well as a further 3. The drawback of signature profiling, however, is that it requires one experiment per deletion strain and thus thousands of arrays to measure all strains, compared to only a single array at each time point for the deletion screen.

When we extended our screen to humans, we found 255 human orthologs of class III yeast genes associated with defects in growth on non-fermentable substrates (see Web Table A online). Of these, 21 were genes known to be involved in mitochondrial disease inherited in a mendelian fashion. These included genes encoding two subunits of complex II of the respiratory chain, five assembly factors of complex III and IV and six proteins associated with deficiencies of mitochondrial multienzyme complexes (Fig. 4). Additionally, eight orthologs were found associated with diseases for which a mitochondrial pathophysiology is plausible but has not been proved. In turn, for 33 of 102 known human genes associated with mendelian mitochondrial disease, there was no corresponding yeast gene whose deletion was associated with growth defects, although in a few cases (including *ALD4*, *GUT2* and *YBR208C*) the deletion strains showed minor deficiencies. A further 15 of the 102 were not measured in our screen (and hence were either lethal or not detectable) and 33 yielded no yeast orthologs (see Web Table B online). Although our systematic screen for human disease genes in yeast was therefore by no means comprehensive, these data showed that many human disease genes are associated with a wide spectrum of yeast deletion phenotypes and can be identified through quantitative growth selection in yeast.

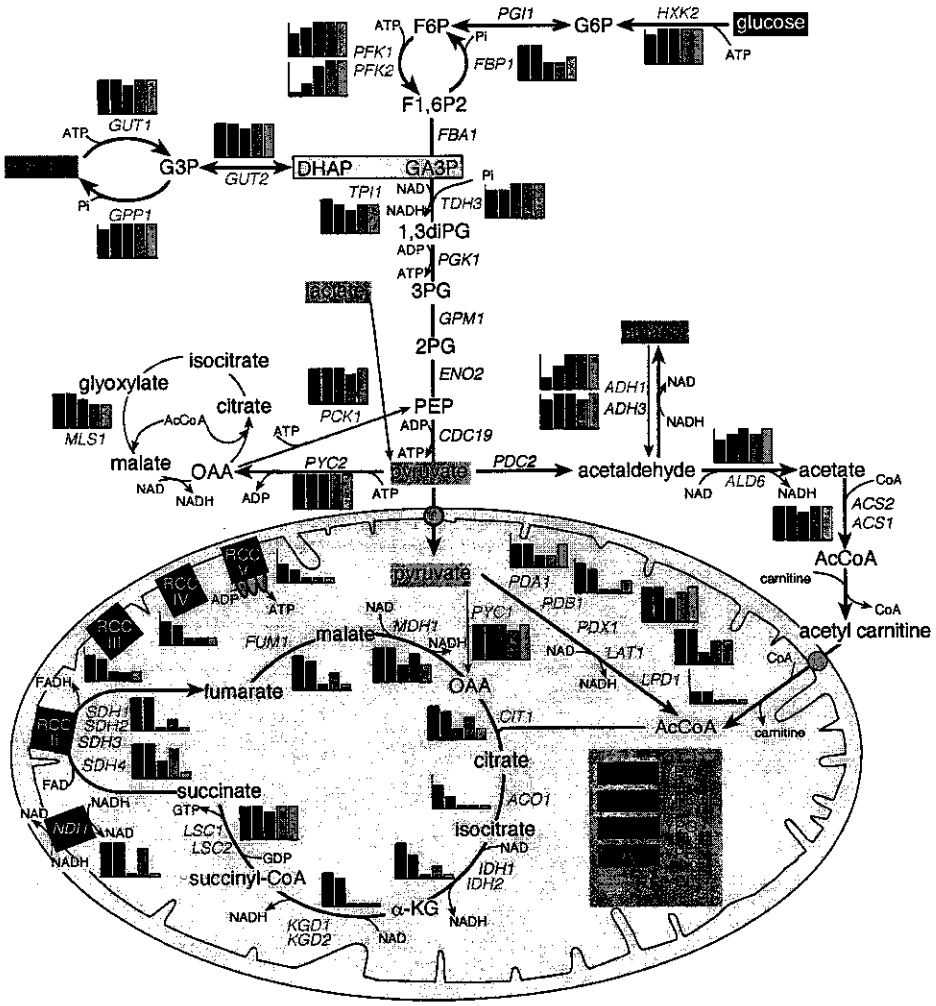


Fig. 3 Distinction between mitochondrial and cytoplasmic pathway branches: glycolysis above and TCA cycle and mitochondrial respiratory chain below. The bar graphs indicate the relative fitness of a homozygous deletion mutant of a gene under different medium conditions, color-coded in the legend. Genes without bar graphs were not detected. For the respiratory-chain complexes (RCC) III, IV and V, an average profile is shown. Deletions of genes shown in red lettering result in deficiency in growth on fermentable substrates, and deletion of those in green in a deficiency in growth on non-fermentable substrates. Ac, acetyl; CoA, coenzyme A; DHAP, dihydroxyacetone phosphate; 1,3diPG, 1,3-bisphosphoglycerate; F6P, fructose-6-phosphate; F1,6P2, fructose-1,6-bisphosphate; GA3P, glyceraldehyde-3-phosphate; α-KG, α-ketoglutarate; OAA, oxaloacetate; PEP, phosphoenolpyruvate; 2PG, 2-phosphoglycerate; 3PG, 3-phosphoglycerate; Pi, phosphate.

To propose specific new disease candidates, we selected seven mapped, putative mitochondrial disorders for which affected individuals have either symptoms characteristic of recorded mitochondrial diseases or biochemical findings indicative of mitochondrial pathophysiology. We analyzed a candidate set of 406 previously known human mitochondrial proteins[10] as well as 259 new proteins identified in our study that were either orthologs to class III proteins or new orthologs to previously known yeast mitochondrial proteins (see Web Table A online). We assigned 24 as candidate genes to reported disease intervals (Table 1). These included 11 new disease candidates identified directly by their class III quantitative deletion phenotype in yeast.

The 6.1-fold enrichment of mitochondrial proteins in the yeast deletion screen exceeded the selection achieved by other systematic, functional, genome-wide approaches. The integration of these data with localization, interaction and other functional information will advance studies of mitochondria in a cellular context. For human diseases, the new genes promise to accelerate positional cloning by serving as candidates for mutational screens in mendelian and complex mitochondrial disorders. Similar approaches may be applicable to other categories of human disease.

strain in the same medium had to be parallel and could therefore be represented by a single regression slope. For easier reading, we added 1.0 to the regression slopes, yielding a value of 1.0 for a lack of change in strain abundance, less than 1.0 for strains growing more slowly than, and more than 1.0 for strains growing more quickly than the average growth rate of the pool.

Clustering to define four phenotypic classes was performed by the *k*-means method, employing the growth estimates (per strain per medium) with the GeneSpring statistical package (Silicon Genetics). We averaged the results of the two duplicate experiments for the homozygous diploid time course in relation to media YPD, YPDGE, YPG, YPE and YPL, and calculated the similarity between genes using standard correlation with a maximum of 100 iterations. Further clustering of genes within a group was carried out using hierarchical clustering.

Calculation of the enrichment factor was carried out using the formula $(a/b)/(c/d)$, where a is number of previously known mitochondrial proteins with a phenotype, b is the total number of genes with a phenotype, c is the number of previously known mitochondrial genes in the genome and d is the total number of genes in the genome. For expression analysis of the diauxic shift[19], we used the pool of genes that had at least either a doubling or a halving of expression ($a = 137$, $b = 1618$, $c = 425$, $d = 5965$, compared with $a = 201$ and $b = 466$ for our data, with spurious ORFs[12] removed from both data sets). Had we considered only genes with at least a doubling of expression, an enrichment factor of 2.2 would have been obtained.

Protein import into isolated mitochondria. For SP6 polymerase–driven synthesis of pre-proteins *in vitro*, we amplified the ORFs from ATG to STOP codon by PCR and cloned into the vector pGEM4 (Promega). Radiolabeled pre-proteins were synthesized by a coupled *in vitro* transcription–translation reaction in reticulocyte lysate (Promega) in the presence of [^{35}S]methionine. After isolating mitochondria from yeast strain W334 grown on lactate medium, we resuspended them at 25 °C in import buffer (0.3 mg/ml fatty acid–free BSA, 0.6 M sorbitol, 80 mM KCl, 10 mM magnesium acetate, 2 mM KH_2PO_4, 2.5 mM EDTA, 2.5 mM $MnCl_2$, 2 mM ATP, 5 mM NADH and 50 mM HEPES–KOH, pH 7.2). We initiated import by adding 1–10% (vol/vol) of reticulocyte lysate containing radiolabeled pre-protein. After 15 min, we placed the samples on ice for 15 min with or without proteinase K (50 μg/ml) to remove non-imported proteins. Protease was inhibited by the addition of 2 mM PMSF. We re-isolated mitochondria and analyzed them by SDS–PAGE and autoradiography. Control experiments were performed in the absence of membrane potential in the presence of 1 μM valinomycin and 20 μM oligomycin.

Human candidate disease gene analysis. We searched yeast ORFs using NCBI tBLASTN against UniGene using as the database Hs.seq.uniq[21] (Oct. 30, 2001 build), containing one sequence selected from each UniGene cluster. Searches were performed with an *E*-value cutoff of 10^{-4} and default values of all other parameters. Genes known to be mutated in human disease were also searched against the yeast genome; differences in the reciprocal BLAST searches were analyzed by hand. Genes known to be involved in human mitochondrial disease were considered only if genetic analysis had confirmed a gene mutation. Ortholog genome positions were identified and searched against diseases listed in the Online Mendelian Inheritance in Man (OMIM) database.

Online supplementary information. Fitness values and growth plots for each yeast deletion strain in each media condition are available online in a searchable database at http://www-deletion.stanford.edu/YDPM/YDPM_index.html. Other information was obtained from the yeast deletion project on http://yeastdeletion.stanford.edu/ and the MitoP database for mitochondrial-related genes, proteins and diseases at http://mips.gsf.de/proj/medgen/mitop/.

GenBank accession numbers. *ACAA2*, NM_006111; *ACO1*, NM_002197; *ALAS2*, NM_000032; *ALDH1B1*, NM_000692; *APEXL2*, BC007669; *ATP5A1*, NM_004046; *CGI-11*, NM_015941; *DKFZP667C165*, XM_042282; *DNAJA1*, NM_001539; *LOC85479*, NM_033105; *MGC14797*, NM_032335; *MGC14836*, NM_033412; *MRPL15*, NM_014175; *NDUFA1*, NM_004541; *NDUFB6*, NM_002493; *PDCD8*, NM_004208; *PDE7A*, XM_037534; *PFKFB1*, NM_002625; *PLS3*, NM_005032; *SLC9A6*, NM_006359; *SLC25A5*, NM_001152; *SLC25A14*, NM_003951; *SR-BP1*, NM_005866; *TIMM17B*, NM_005834.

Note: Supplementary information is available on the Nature Genetics website.

Acknowledgments

We thank M. Mindrinos, E. Allen, T. Neklesa, Q. Wang, W. Neupert and T. Meitinger for helpful advice and M. Trebo for help with preparing the supplementary website. This work was supported by the US National Institutes of Health (P.J.O. and R.W.D.) and the Bundesministerium für Bildung und Forschung (H.P.). L.M.S. was supported as a Howard Hughes Medical Institute predoctoral fellow and C.S. as a Deutsche Forschungsgemeinschaft postdoctoral fellow.

Competing interests statement

The authors declare that they have no competing financial interests.

Received 20 February; accepted 16 May 2002.

1. Foury, F. Human genetic diseases: a cross-talk between man and yeast. *Gene* **195**, 1–10 (1997).
2. DiMauro, S. & Schon, E.A. Nuclear power and mitochondrial disease. *Nature Genet.* **19**, 214–215 (1998).
3. Wallace, D.C. Mitochondrial diseases in man and mouse. *Science* **283**, 1482–1488 (1999).
4. Winzeler, E.A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
5. Birrell, G.W., Giaever, G., Chu, A.M., Davis, R.W. & Brown, J.M. A genome-wide screen in *Saccharomyces cerevisiae* for genes affecting UV radiation sensitivity. *Proc. Natl Acad. Sci. USA* **98**, 12608–12613 (2001).
6. Ooi, S.L., Shoemaker, D.D. & Boeke, J.D. A DNA microarray-based genetic screen for nonhomologous end-joining mutants in *Saccharomyces cerevisiae*. *Science* **294**, 2552–2556 (2001).
7. Shoemaker, D.D., Lashkari, D.A., Morris, D., Mittmann, M. & Davis, R.W. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nature Genet.* **14**, 450–456 (1996).
8. Tzagoloff, A. & Myers, A.M. Genetics of mitochondrial biogenesis. *Annu. Rev. Biochem.* **55**, 249–285 (1986).
9. Yaffe, M.P. in *Methods in Enzymology* Vol. 194 (eds Guthrie, C. & Fink, G.R.) 627–643 (Academic Press, San Diego, California, 1991).
10. Scharfe, C. *et al.* MITOP, the mitochondrial proteome database: 2000 update. *Nucleic Acids Res.* **28**, 155–158 (2000).
11. Grivell, L.A. *et al.* Mitochondrial assembly in yeast. *FEBS Lett.* **452**, 57–60 (1999).
12. Wood, V., Rutherford, K.M., Ivens, A., Rajandream, M.A. & Barrell, B. A re-annotation of the *Saccharomyces cerevisiae* genome. *Comp. Funct. Genom.* **2**, 143–154 (2001).
13. Claros, M.G. & Vincens, P. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.* **241**, 779–786 (1996).
14. Andersson, S.G. *et al.* The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**, 133–140 (1998).
15. Schwikowski, B., Uetz, P. & Fields, S. A network of protein–protein interactions in yeast. *Nature Biotechnol.* **18**, 1257–1261 (2000).
16. Kumar, A. *et al.* Subcellular localization of the yeast proteome. *Genes Dev.* **16**, 707–719 (2002).
17. Nishino, I., Spinazzola, A. & Hirano, M. Thymidine phosphorylase gene mutations in MNGIE, a human mitochondrial disorder. *Science* **283**, 689–692 (1999).
18. Ashe, M.P., De Long, S.K. & Sachs, A.B. Glucose depletion rapidly inhibits translation initiation in yeast. *Mol. Biol. Cell* **11**, 833–848 (2000).
19. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
20. Hughes, T.R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
21. Boguski, M.S. & Schuler, G.D. ESTablishing a human transcript map. *Nature Genet.* **10**, 369–371 (1995).
22. Mewes, H.W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **28**, 37–40 (2000).
23. Hentati, A. *et al.* Linkage of 'pure' autosomal recessive familial spastic paraplegia to chromosome 8 markers and evidence of genetic locus heterogeneity. *Hum. Mol. Genet.* **3**, 1263–1267 (1994).
24. Christodoulou, K. *et al.* Mapping of the second Friedreich's ataxia (FRDA2) locus to chromosome 9p23–p11: evidence for further locus heterogeneity. *Neurogenetics* **3**, 127–132 (2001).
25. Kerrison, J.B. *et al.* Genetic heterogeneity of dominant optic atrophy, Kjer type: identification of a second locus on chromosome 18q12.2–12.3. *Arch. Ophthalmol.* **117**, 805–810 (1999).
26. Assink, J.J. *et al.* A gene for X-linked optic atrophy is closely linked to the Xp11.4–Xp11.2 region of the X chromosome. *Am. J. Hum. Genet.* **61**, 934–939 (1997).
27. Priest, J.M., Fischbeck, K.H., Nouri, N. & Keats, B.J. A locus for axonal motor-sensory neuropathy with deafness and mental retardation maps to Xq24–q26. *Genomics* **29**, 409–412 (1995).
28. McMullan, T.F., Collins, A.R., Tyers, A.G. & Robinson, D.O. A novel X-linked dominant condition: X-linked congenital isolated ptosis. *Am. J. Hum. Genet.* **66**, 1455–1460 (2000).
29. Malmgren, H. *et al.* Linkage mapping of a severe X-linked mental retardation syndrome. *Am. J. Hum. Genet.* **52**, 1046–1052 (1993).