La génomique es
Elle travaille au sé
aux différents gèn

La **protéomique**
c'est à dire l'enser
(leur rôle, leur stru
organisme , un tis

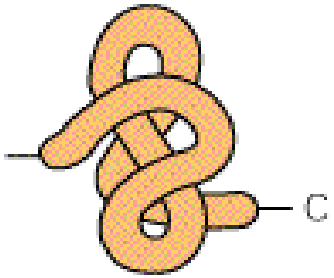
La métabolomiqu
étudie l'ensemble
présents dans un

DNA

Transcription

RNA

Translation



Protein

Substrat → Produits

Génomique

Transcriptomique

Protéomique

Métabolomique

CONNAITRE LE GENOME

La génomique

Structure physique

Structure génétique

Les espèces modèles

Plantes

Champignons

Procaryotes

Animaux

↓

Un ancêtre commun ?
LUCA

Last Universal Common Ancestor

Les objectifs

- Localiser et séquencer les gènes

- Étudier la fonction des gènes

Analyse fonctionnelle

Les applications

Construction de génotypes élites

Par :

- la fourniture d'une grande quantité de marqueurs moléculaires
- un meilleur contrôle de la régulation des gènes
- l'identification de gènes candidats pour l'analyse des QTL de caractères majeurs
- l'identification d'allèles favorables

La Biologie moléculaire

et

le développement de la bioinformatique

La Biologie moléculaire (mi 80- début 1990)

Le clonage et le séquençage d'ADN deviennent progressivement pratiques courantes.

L'utilisation de l'outil informatique est devenu indispensable pour réaliser les expériences de Biologie Moléculaire et pour en analyser les résultats

→ Au départ, c'est la BM qui a besoin de la Bioinfo

Internet, BLAST et séquencage

Internet : L'accès aux bases de données et aux outils d'analyse devient transparent et facile pour tous

BLAST : rechercher dans les bases de données l'existence de séquences similaires à une séquence donnée c'est la fonction utilisée dans plus de 90% des analyses. BLAST (1990) permet à tout un chacun de le faire, en classant les résultats significatifs

Le séquençage génomique : il donne accès d'un seul coup à toute l'information génétique d'un organisme … mais comment la gérer, la déchiffrer et l'utiliser ?

explosion de nouveaux besoins pour la bioinformatique
explosion de données d'un type nouveau à exploiter

Besoins

Gérer les programmes de séquençage

Annoter les génomes : où sont les gènes? Que font-ils ?

Etablir des ontologies et les peupler : de quoi parle t'on au juste ?

Permettre la comparaison de génomes entre eux

BM → ← Bioinfo

→ Si au départ, c'est la BM qui avait besoin de la Bioinfo

→ aujourd'hui, la bioinfo a aussi besoin des données générées par la BM

CONNAITRE LE GENOME

La génomique

Structure physique **Structure génétique**

Les espèces modèles

Quels modèles choisir ?

Plantes

Champignons

Procaryotes

Animaux

Les objectifs

- Localiser et séquencer les gènes

- Étudier la fonction des gènes

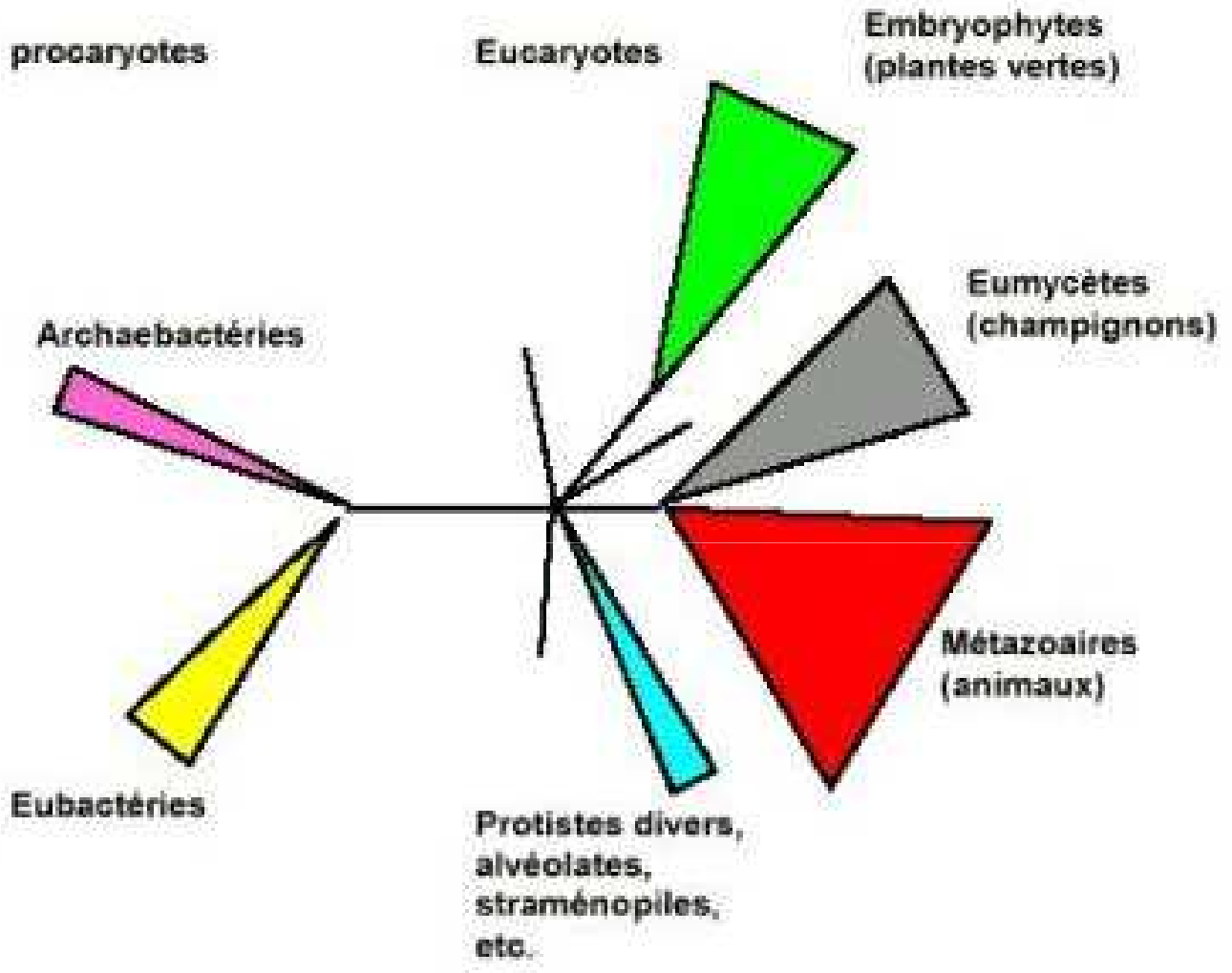
Les applications

Construction de génotypes élites

Par :

- la fourniture d'une grande quantité de marqueurs moléculaires
- un meilleur contrôle de la régulation des gènes
- l'identification de gènes candidats pour l'analyse des QTL de caractères majeurs
- l'identification d'allèles favorables

Analyse fonctionnelle



Les triangles représentent l'effectif approximatif en espèces de chaque groupe

Le choix des espèces modèles se fera sur la taile desgenomes et
sur leur facilite de « culture » en laboratoire

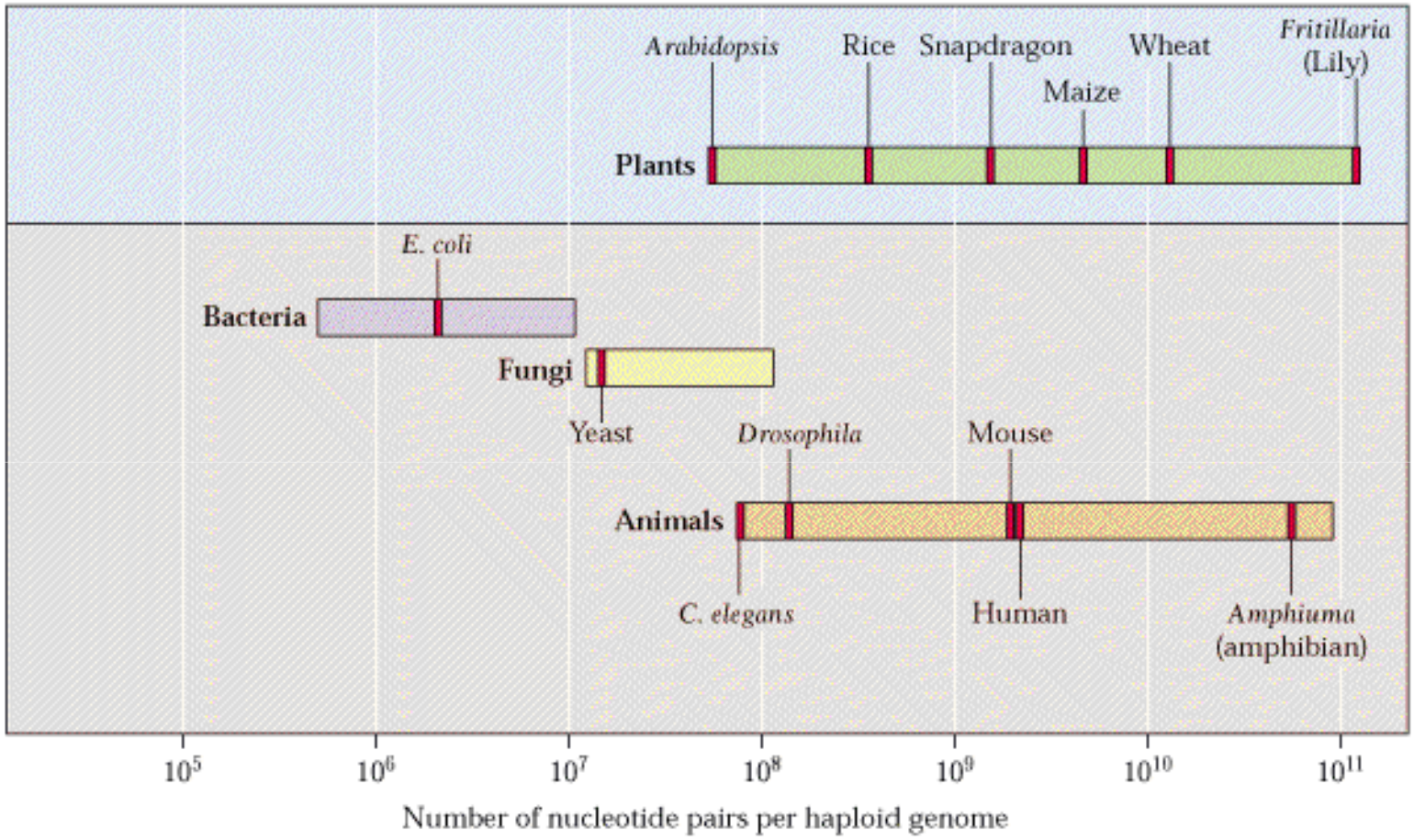
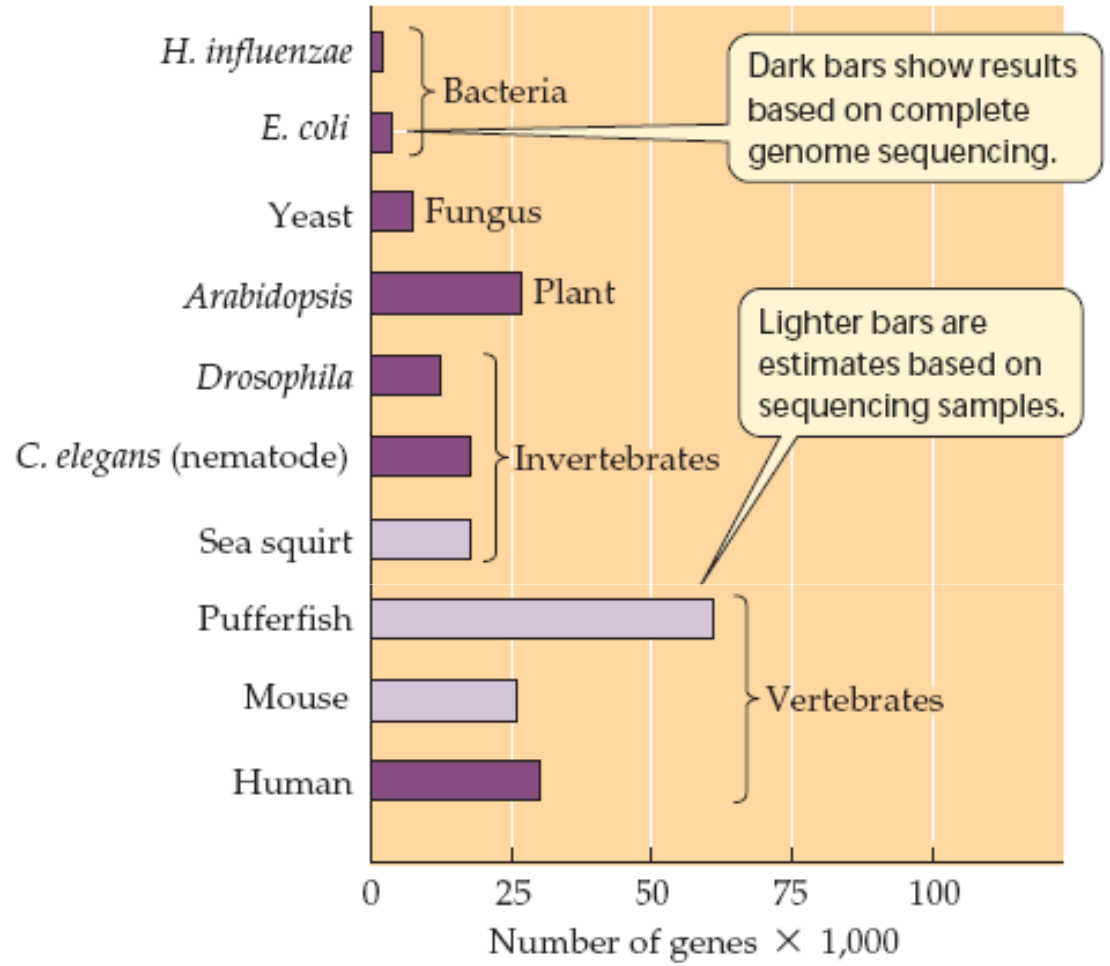


Figure 7.19
C values (haploid genome size in basepairs) from various organisms.

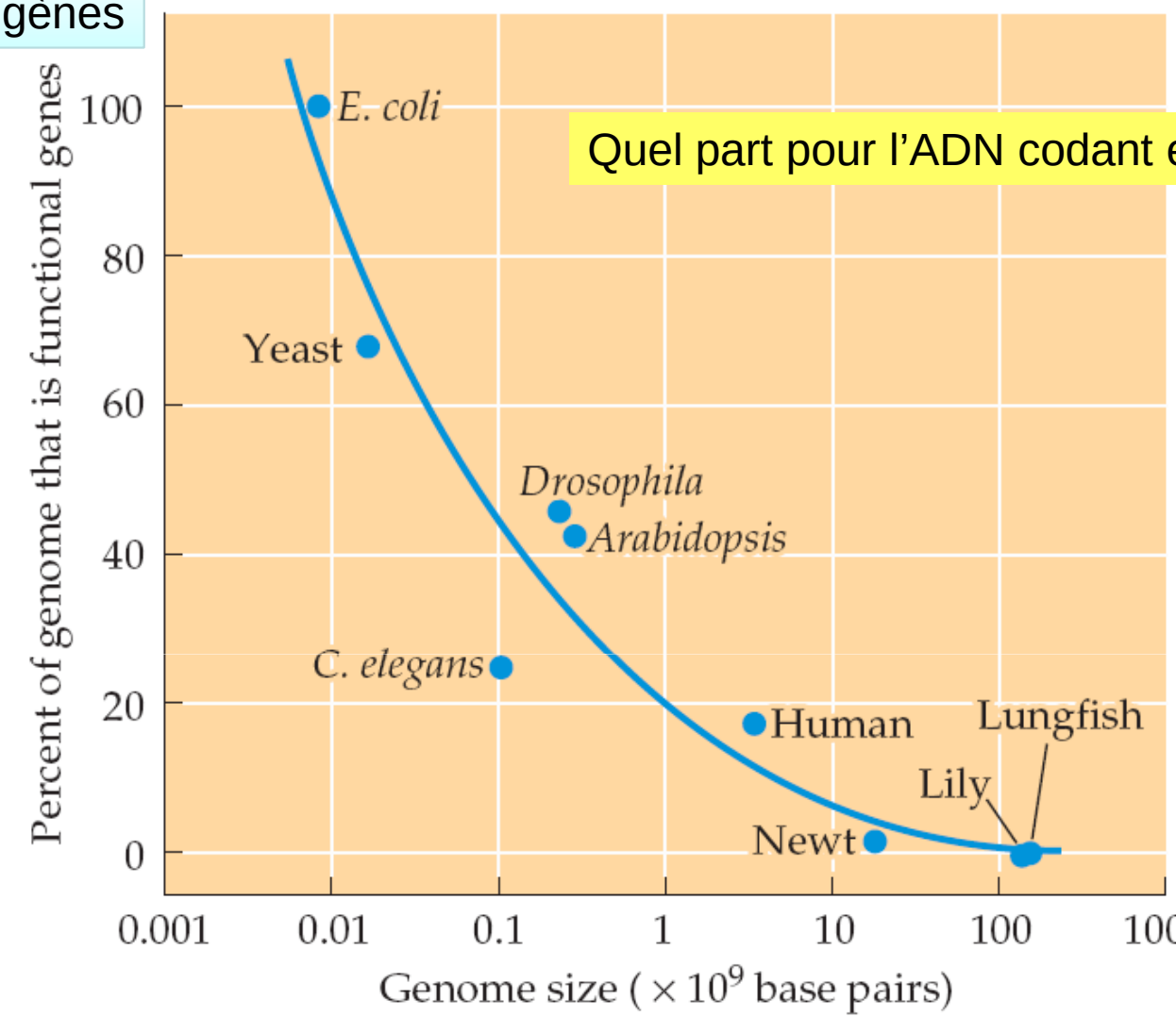
Print Bac

Combien de gènes par organismes ?



26.7 Complex Organisms Have More Genes than Simpler Organisms Genome sizes have been measured or estimated in a variety of organisms ranging from single-celled prokaryotes to vertebrates.

gènes



Quel part pour l'ADN codant et non–codant ?

Taille du génome

26.8 A Large Proportion of DNA Is Noncoding Most of the DNA of bacteria and yeasts encodes RNAs or proteins, but most of the DNA of more complex organisms is noncoding. Most noncoding DNA is probably nonfunctional.

Les programmes de séquençage du génome

- Les genomes center ?
- Les programmes ESTs
- Le séquençage des génomes
- L’annotation des génomes
- Les centres de ressources

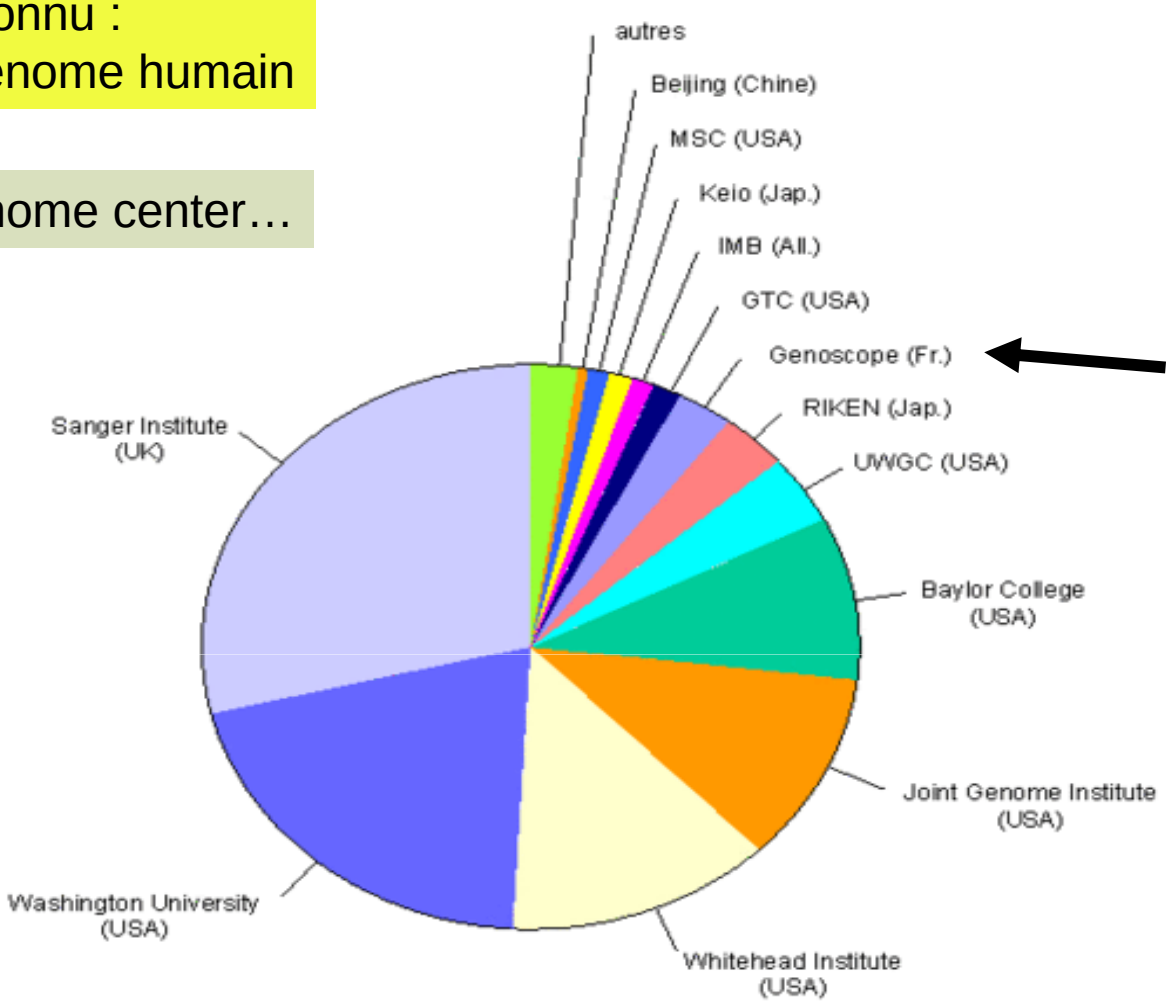
Les genomes centers
Human Genome Sequencing Center / Baylor College of Medicine, Houston (Texas) ; USA
Human Genome Center / Beijing Genomics Institute, Académie chinoise des sciences, Beijing ; Chine
Lita Annenberg Hazen Genome Center / Cold Spring Harbor Laboratory, Cold Spring Harbor (N.Y.), USA
Gesellschaft fur Biotechnologische Forschung mbH, Braunschweig ; Allemagne
Genoscope, Evry ; France
GTC Sequencing Center / Genome Therapeutics Corp., Waltham (Mass.) ; USA
Department of Genome Analysis / Institute of Molecular Biotechnology, Jena ; Allemagne
Joint Genome Institute / U.S. Department of Energy, Walnut Creek (Calif.) ; USA
Département de biologie moléculaire / Ecole de médecine de l'université Keio, Tokyo ; Japon
Max Planck Institute for Molecular Genetics, Berlin ; Allemagne
Multimegabase Sequencing Center / The Institute for Systems Biology, Seattle (Wash.) ; USA
RIKEN Genomic Sciences Center, Yokohama ; Japon
The Wellcome Trust Sanger Institute (Sanger Center), Hinxton ; Royaume-Uni
Stanford Human Genome Center, Stanford (Calif.) ; USA
University of Oklahoma / Advanced Center for Genome Technology, Norman (Okla.), USA
Whitehead Institute / MIT Center for Genome Research, Cambridge (Mass.) ; USA
Washington University / Genome Sequencing Center, St Louis (Mo.) ; USA

USA

Les capacites de séquençage sont exponentielles : et chaque centre ou federation de labos a ses propres sequenceurs → Vers du sequençage de paillasse

L’exemple le mieux connu :
Le séquençage du génome humain

Par genome center…



Sur le plan international, les contributions des 6 pays impliqués dans le projet sont les suivantes :

Par pays …

Etats-Unis	60,8 %
Royaume-Uni	28,9 %
Japon	4,9 %
France	2,8 %
Allemagne	1,5 %
Chine	0,7 %

Le séquençage du génome humain

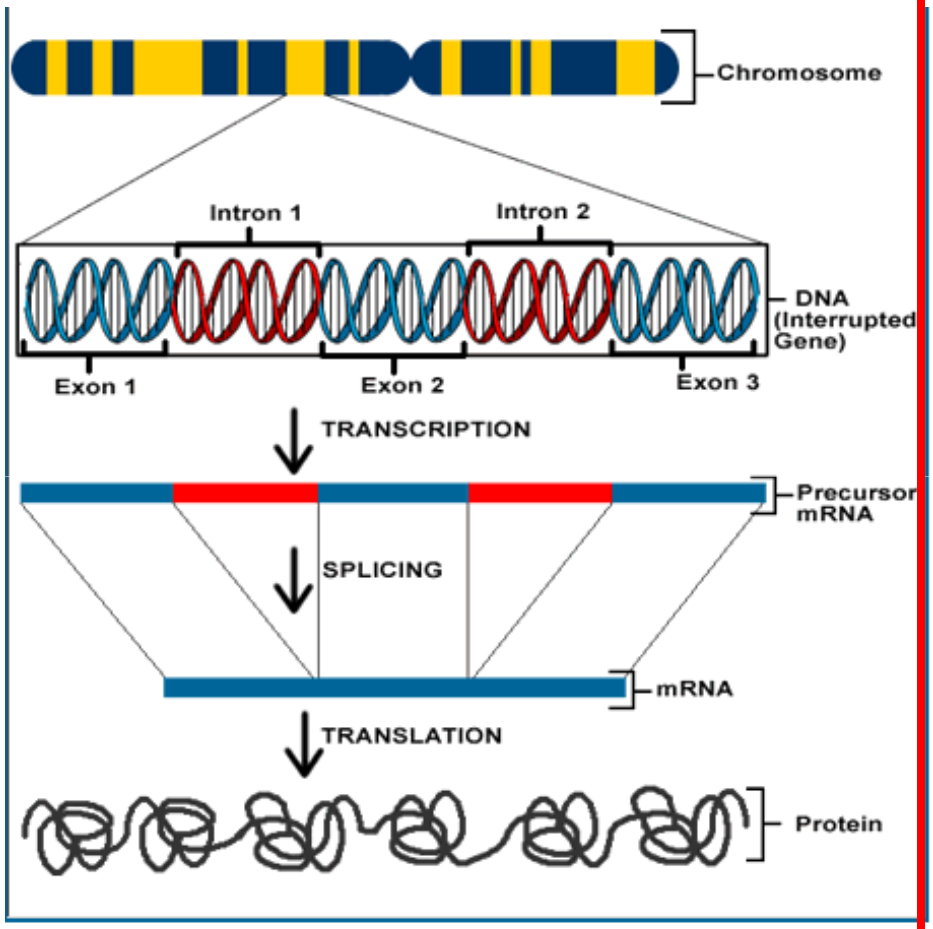
Le coût total du projet Génome humain est d'environ 2,7 milliards de dollars alors qu'il avait été estimé à 3 milliards de dollars au début du projet, en 1990.

Cette économie résulte de **progrès techniques considérables et de l'accélération du projet, terminé avec deux ans d'avance sur les prévisions.**

Une grande part de la somme a été dépensée pour la finition de l'ébauche génomique obtenue en 2000.

Le séquençage du chromosome 14 a coûté environ 10 millions d'euros,
+
plusieurs millions d'euros pour l'analyse et l'annotation.

ESTs



igure 1. An overview of the process of protein synthesis.

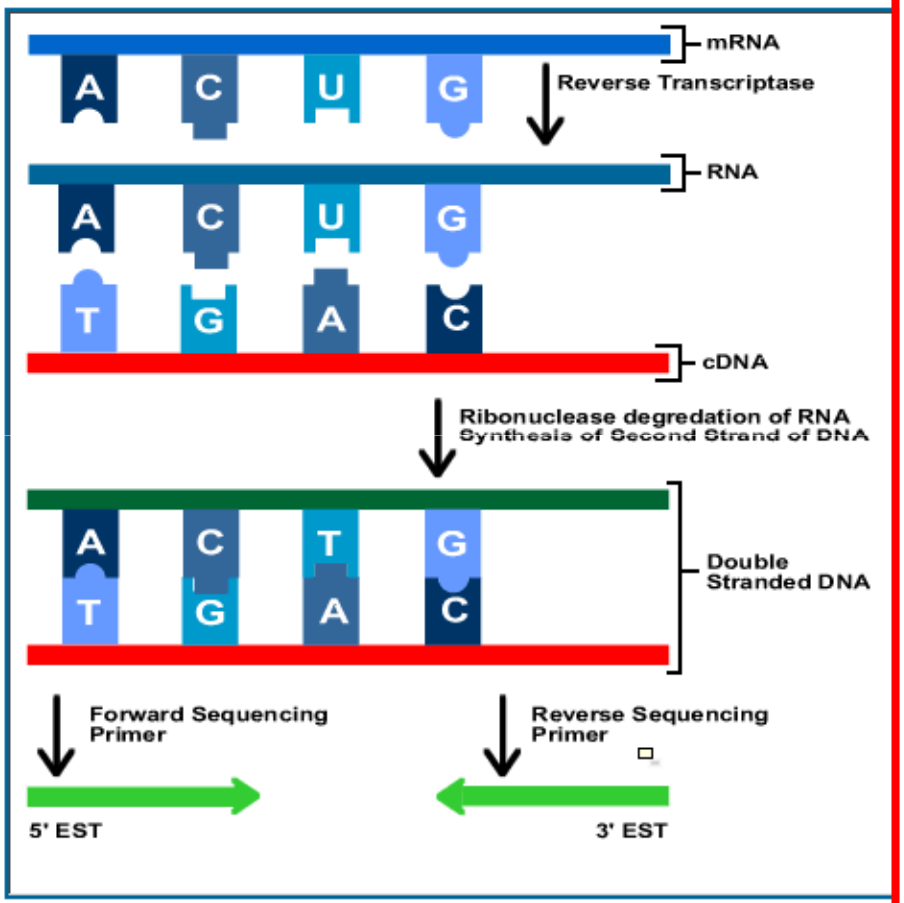


Figure 2. An overview of how ESTs are generated.

Banques ADNc (multiples conditions cultures …)

→Clone pris au hasard, séquencage et dépôt de la séquence dans une base de donnée et du clone dans un centre de ressource génétique

Séquençage (plutot en 5') de 300-800n en 1 seule réaction

Summary by Organism - Jan 2, 1997

Number of public entries: 775,836

Homo sapiens (human)	539,706
Mus musculus + domesticus (mouse)	120,098
Caenorhabditis elegans (nematode)	30,196
Arabidopsis thaliana (thale cress)	29,166
Oryza sativa (rice)	12,672
Toxoplasma gondii	8,165
Rattus sp. (rat)	6,615
Brugia malayi (parasitic nematode)	5,184
Saccharomyces cerevisiae (baker's yeast)	2,944
Caenorhabditis briggsae	2,424
Trypanosoma brucei rhodesiense	2,175
Plasmodium falciparum (malaria parasite)	2,011
Schistosoma mansoni (blood fluke)	1,958
Zea mays (maize)	1,757
Drosophila melanogaster (fruit fly)	1,690
Brassica napus (oilseed rape)	1,427
Sus scrofa (pig)	1,055
Brassica campestris (field mustard)	965
Ricinus communis (castor bean)	750
Pyrococcus furiosus (hyperthermophilic archaeon)	577
Leishmania major	527
Citrus unshiu	426
Onchocerca volvulus	306
Emericella nidulans	272
Pinus taeda (loblolly pine)	256
Capra hircus (goat)	245
Strongylocentrotus purpuratus (purple urchin)	242
Oryctolagus cuniculus (rabbit)	206
Gallus gallus (chicken)	191
Danio rerio (zebrafish)	128
Schistosoma japonicum (blood fluke)	109
Saccharum sp.	105
Wuchereria bancrofti	99
Glycine max (soybean)	97
Sorghum bicolor (sorghum)	91
Citrus sinensis	87
Aspergillus niger	85
Pristionchus pacificus	78
Cryptosporidium parvum	65
Bos taurus (cattle)	60
Strongyloides stercoralis	57
Trypanosoma brucei brucei	56
Solanum tuberosum (potato)	49
Filobasidiella neoformans (Cryptococcus neoformans)	44
Macropus eugenii (tammar wallaby)	42

Summary by Organism - January 21, 2000

Number of public entries: 3,483,128

Homo sapiens (human)	1,715,552
Mus musculus + domesticus (mouse)	904,535
Rattus sp. (rat)	132,680
Caenorhabditis elegans (nematode)	101,232
Drosophila melanogaster (fruit fly)	86,121
Danio rerio (zebrafish)	54,600
Lycopersicon esculentum (tomato)	50,303
Oryza sativa (rice)	47,474
Zea mays (maize)	47,231
Arabidopsis thaliana (thale cress)	45,757
Glycine max (soybean)	40,046
Brugia malayi (parasitic nematode)	20,773
Dictyostelium discoideum	19,182
Emericella nidulans	12,993
Schistosoma mansoni (blood fluke)	12,509
Chlamydomonas reinhardtii	11,690
Toxoplasma gondii	10,741
Neurospora crassa	10,229
Pinus taeda (loblolly pine)	9,870
Gossypium hirsutum (upland cotton)	9,335
Trypanosoma cruzi	8,796
Onchocerca volvulus	8,674
Schizosaccharomyces pombe (fission yeast)	8,118
Bombyx mori (domestic silkworm)	6,508
Medicago truncatula (barrel medic)	6,059
Xenopus laevis (African clawed frog)	5,853
Sorghum bicolor (sorghum)	5,428
Sus scrofa (pig)	5,194
Pristionchus pacificus	4,989
Populus tremula x Populus tremuloides	4,809
Trypanosoma brucei rhodesiense	4,742
Mesembryanthemum crystallinum (common ice plant)	4,429
Gallus gallus (chicken)	4,194
Halocynthia roretzi	4,163
Eimeria tenella	3,721
Saccharomyces cerevisiae (baker's yeast)	3,041
Plasmodium falciparum (malaria parasite)	2,871
Caenorhabditis briggsae	2,424
Leishmania major	2,191
Oryctolagus cuniculus (rabbit)	2,094
Physcomitrella patens	1,909
Paralichthys olivaceus	1,787
Brassica napus (oilseed rape)	1,702
Metarhizium anisopliae	1,693
Ceratodon purpureus	1,663
Bos taurus (cattle)	1,610
Schistosoma japonicum (blood fluke)	1,395
Aedes aegypti (yellow fever mosquito)	1,327

X4.5

Summary by Organism - October 13, 2000

Number of public entries: 6,051,828

Homo sapiens (human)	2,519,935
Mus musculus + domesticus (mouse)	1,703,709
Rattus sp. (rat)	188,794
Bos taurus (cattle)	134,604
Glycine max (soybean)	122,843
Arabidopsis thaliana (thale cress)	112,467
Caenorhabditis elegans (nematode)	101,252
Drosophila melanogaster (fruit fly)	95,211
Lycopersicon esculentum (tomato)	87,680
Medicago truncatula (barrel medic)	81,694
Zea mays (maize)	73,965
Danio rerio (zebrafish)	73,703
Oryza sativa (rice)	62,798
Chlamydomonas reinhardtii	55,860
Sorghum bicolor (sorghum)	45,265
Triticum aestivum (wheat)	44,132
Xenopus laevis (African clawed frog)	43,935
Sus scrofa (pig)	33,465
Lotus japonicus	26,844
Neurospora crassa	24,635
Hordeum vulgare (barley)	23,519
Brugia malayi (parasitic nematode)	22,392
Pinus taeda (loblolly pine)	22,030
Dictyostelium discoideum	19,183
Solanum tuberosum (potato)	14,043
Bombyx mori (domestic silkworm)	14,807
Gossypium arboreum	13,662
Emericella nidulans	12,993
Schistosoma mansoni (blood fluke)	12,959
Gallus gallus (chicken)	12,840
Onchocerca volvulus	12,552
Toxoplasma gondii	12,177
Mesembryanthemum crystallinum (common ice plant)	11,115
Ciona intestinalis	10,347
Porphyra yezoensis	10,185
Trypanosoma cruzi	9,919
Gossypium hirsutum (upland cotton)	9,438
Schizosaccharomyces pombe (fission yeast)	8,118
Physcomitrella patens	7,496
Strongyloides stercoralis	7,343
Meloidogyne incognita (southern root-knot nematode)	6,626
Secale cereale	6,574
Anopheles gambiae (African malaria mosquito)	6,023
Eimeria tenella	5,499
Pristionchus pacificus	4,989
Trypanosoma brucei rhodesiense	4,821
Populus tremula x Populus tremuloides	4,809
Halocynthia roretzi	4,163

X 1.7

Summary by Organism - September 26, 2003

Number of public entries: 18,666,654

Homo sapiens (human)	5,425,176
Mus musculus + domesticus (mouse)	3,874,479
Rattus sp. (rat)	537,607
Triticum aestivum (wheat)	499,979
Ciona intestinalis	492,488
Gallus gallus (chicken)	451,109
Zea mays (maize)	362,549
Danio rerio (zebrafish)	362,279
Hordeum vulgare + subsp. vulgare (barley)	348,225
Glycine max (soybean)	341,573
Bos taurus (cattle)	319,805
Xenopus laevis (African clawed frog)	311,905
Drosophila melanogaster (fruit fly)	261,404
Oryza sativa (rice)	260,879
Saccharum officinarum	246,301
Caenorhabditis elegans (nematode)	215,200
Silurana tropicalis	203,883
Arabidopsis thaliana (thale cress)	188,782
Medicago truncatula (barrel medic)	187,763
Dictyostelium discoideum	155,032
Chlamydomonas reinhardtii	154,600
Sorghum bicolor (sorghum)	151,874
Lycopersicon esculentum (tomato)	150,228
Sus scrofa (pig)	150,133
Schistosoma mansoni (blood fluke)	139,064
Anopheles gambiae (African malaria mosquito)	130,731
Vitis vinifera	128,211
Oryzias latipes (Japanese medaka)	103,098
Oncorhynchus mykiss (rainbow trout)	102,218
Pinus taeda (loblolly pine)	100,541
Solanum tuberosum (potato)	94,473
Toxoplasma gondii	72,859
Lactuca sativa	68,188
Helianthus annuus	59,837
Salmo salar	58,330
Populus tremula x Populus tremuloides	56,013
Strongylocentrotus purpuratus (purple urchin)	51,744
Physcomitrella patens subsp. patens	49,583
Schistosoma japonicum (blood fluke)	45,902
Ascaris suum (pig roundworm)	39,242
Gossypium arboreum	38,915
Brassica napus (oilseed rape)	37,108
Lotus corniculatus var. japonicus	36,262
Hydra magnipapillata	35,154
Magnaporthe grisea	31,397
Bombyx mori (domestic silkworm)	28,978
Eimeria tenella	28,550
Neurospora crassa	28,089
Canis familiaris (dog)	27,010

X 3

http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html

dbEST: database of "Expressed Sequence Tags"

dbEST release 091908

Summary by Organism - September 19, 2008

X 3

Number of public entries: 55,785,287

Homo sapiens (human)	8,138,094
Mus musculus + domesticus (mouse)	4,850,258
Arabidopsis thaliana (thale cress)	1,526,133
Bos taurus (cattle)	1,517,053
Sus scrofa (pig)	1,476,546
Zea mays (maize)	1,464,859
Danio rerio (zebrafish)	1,379,829
Xenopus (Silurana) tropicalis (western clawed frog)	1,271,375
Oryza sativa (rice)	1,220,876
Ciona intestinalis	1,204,893
Triticum aestivum (wheat)	1,051,300
Rattus norvegicus + sp. (rat)	895,894
Glycine max (soybean)	839,041
Xenopus laevis (African clawed frog)	677,784
Oryzias latipes (Japanese medaka)	616,739
Gallus gallus (chicken)	599,610
Brassica napus (oilseed rape)	596,249
Drosophila melanogaster (fruit fly)	573,749
Hordeum vulgare + subsp. vulgare (barley)	478,734
Panicum virgatum (switchgrass)	436,535
Salmo salar (Atlantic salmon)	433,337
Canis lupus familiaris (dog)	365,909
Vitis vinifera (wine grape)	352,984
Caenorhabditis elegans (nematode)	346,109
Branchiostoma floridae (Florida lancelet)	334,502
Pinus taeda (loblolly pine)	328,628
Physcomitrella patens subsp. patens	305,606
Aedes aegypti (yellow fever mosquito)	301,342
Ictalurus punctatus (channel catfish)	300,678
Gasterosteus aculeatus (three spined stickleback)	276,992

NCBI

dbEST: database of "Expressed Sequen Tags"

dbEST release 091109

Summary by Organism - September 11, 2009

Number of public entries: 62,950,194

+ ~ 8 000 000 /année 2008

Homo sapiens (human)	8,296,272
Mus musculus + domesticus (mouse)	4,852,144
Zea mays (maize)	2,018,634
Sus scrofa (pig)	1,536,375
Arabidopsis thaliana (thale cress)	1,527,298
Bos taurus (cattle)	1,517,145
Danio rerio (zebrafish)	1,481,930
Glycine max (soybean)	1,422,497
Xenopus (Silurana) tropicalis (western clawed frog)	1,271,375
Oryza sativa (rice)	1,248,955
Ciona intestinalis	1,205,674
Triticum aestivum (wheat)	1,067,115
Rattus norvegicus + sp. (rat)	1,009,817
Drosophila melanogaster (fruit fly)	821,005
Xenopus laevis (African clawed frog)	677,806
Oryzias latipes (Japanese medaka)	665,382
Brassica napus (oilseed rape)	631,983
Gallus gallus (chicken)	600,075
Hordeum vulgare + subsp. vulgare (barley)	501,614
Salmo salar (Atlantic salmon)	494,152
Panicum virgatum (switchgrass)	436,535
Phaseolus coccineus	391,138
Canis lupus familiaris (dog)	365,909
Physcomitrella patens subsp. patens	362,131
Vitis vinifera (wine grape)	357,849
Caenorhabditis elegans (nematode)	354,744
Ictalurus punctatus (channel catfish)	354,434
Branchiostoma floridae (Florida lancelet)	334,502
Pinus taeda (loblolly pine)	328,628
Malus x domestica (apple tree)	324,308
Ovis aries (sheep)	323,866

NCBI

dbEST: database of "Expressed Sequence Tags"

dbEST release 082710

Summary by Organism - August 27, 2010

Number of public entries: 66,682,428

Homo sapiens (human)	8,304,690
Mus musculus + domesticus (mouse)	4,852,147
Zea mays (maize)	2,019,105
Sus scrofa (pig)	1,621,000
Bos taurus (cattle)	1,559,485
Arabidopsis thaliana (thale cress)	1,529,555
Danio rerio (zebrafish)	1,481,936
Glycine max (soybean)	1,459,936
Xenopus (Silurana) tropicalis (western clawed frog)	1,271,375
Oryza sativa (rice)	1,249,124
Ciona intestinalis	1,205,674
Rattus norvegicus + sp. (rat)	1,162,136
Triticum aestivum (wheat)	1,071,199
Drosophila melanogaster (fruit fly)	821,005
Xenopus laevis (African clawed frog)	677,806
Oryzias latipes (Japanese medaka)	665,382
Brassica napus (oilseed rape)	643,884
Gallus gallus (chicken)	600,418
Panicum virgatum (switchgrass)	546,245
Hordeum vulgare + subsp. vulgare (barley)	501,620
Salmo salar (Atlantic salmon)	496,300
Caenorhabditis elegans (nematode)	393,714
Phaseolus coccineus	391,150
Canis lupus familiaris (dog)	382,618
Vitis vinifera (wine grape)	362,193
Physcomitrella patens subsp. patens	362,131
Ictalurus punctatus (channel catfish)	354,466
Ovis aries (sheep)	335,950
Branchiostoma floridae (Florida lancelet)	334,502
Pinus taeda (loblolly pine)	328,628

dbEST release 120701
Summary by Organism
- 01 July 2012

Number of public entries: 73,360,923

X 100 en 15 ans

The Plant Journal
Volume 4 Page 1051 - December 1993
doi:10.1046/j.1365-313X.1993.04061051.x
Volume 4 Issue 6

An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from *Arabidopsis thaliana*†

Herman Höfte, Thierry Desprez, Joëlle Amselem, Hélène Chiapello and Michel Caboche, Annick Moisan, Marie-Françoise Jourjon and Jean-Louis Charpenteau, Pierre Berthomieu, Danièle Guerrier and Jérome Giraudat, Françoise Quigley, Frank Thomas, De-Yao Yu and Régis Mache, Monique Raynal, Richard Cooke, Françoise Grellet and Michel Delseny, Yves Parmentier, Guy de Marcillac, Claude Gigot, Jacqueline Fleck and Gabriel Philipps, Michèle Axelos, Claude Bardet, Dominique Tremousaygue and Bernard Lescure*

As part of the goal to generate a detailed transcript map for *Arabidopsis thaliana*, 1152 single run sequences (expressed sequence tags or ESTs) have been determined from cDNA clones taken at random in libraries prepared from different sources of plant material: developing siliques, etiolated seedlings, flower buds, and cultured cells. Eight hundred and ninety-five different genes could be identified, 32% of which showed significant similarity to existing sequences in *Arabidopsis* and an array of

Plant Physiol. (1994) 106: 1241–1255

Genes Galore: A Summary of Methods for Accessing Results from Large-Scale Partial Sequencing of Anonymous *Arabidopsis* cDNA Clones[1]

Tom Newman, Frans J. de Bruijn, Pam Green, Ken Keegstra, Hans Kende, Lee McIntosh, John Ohlrogge, Natasha Raikhel, Shauna Somerville, Mike Thomashow, Ernie Retzel, and Chris Somerville*

Arabidopsis Expressed Sequence Tag Project, Department of Energy Plant Research Laboratory, Michigan State University, East Lansing, Michigan 48824 (T.N., F.J.d.B., P.G., K.K., H.K., L.M., J.O., N.R., M.T.); Computational Biology Center, Medical School, University of Minnesota, 1460 Mayo, UMHC 196, 420 Delaware Street S.E., Minneapolis, Minnesota 55455–0312 (E.R.); and Carnegie Institution of Washington, Department of Plant Biology, 290 Panama Street, Stanford, California 94305–4101 (S.S., C.S.)

En 1990, on pouvait publier des séquences ESTs !!

Volume 9 Issue 1

☐ **1:** FEBS Lett. 1997 Mar 24;405(2):129-32.

The Arabidopsis thaliana cDNA sequencing projects.

Delseny M, Cooke R, Raynal M, Grellet F.

University of Perpignan, Laboratoire de Physiologie et Biologie, Moleculaire des Plantes, UMR 5545 CNRS, France. delseny@univ-perp.fr

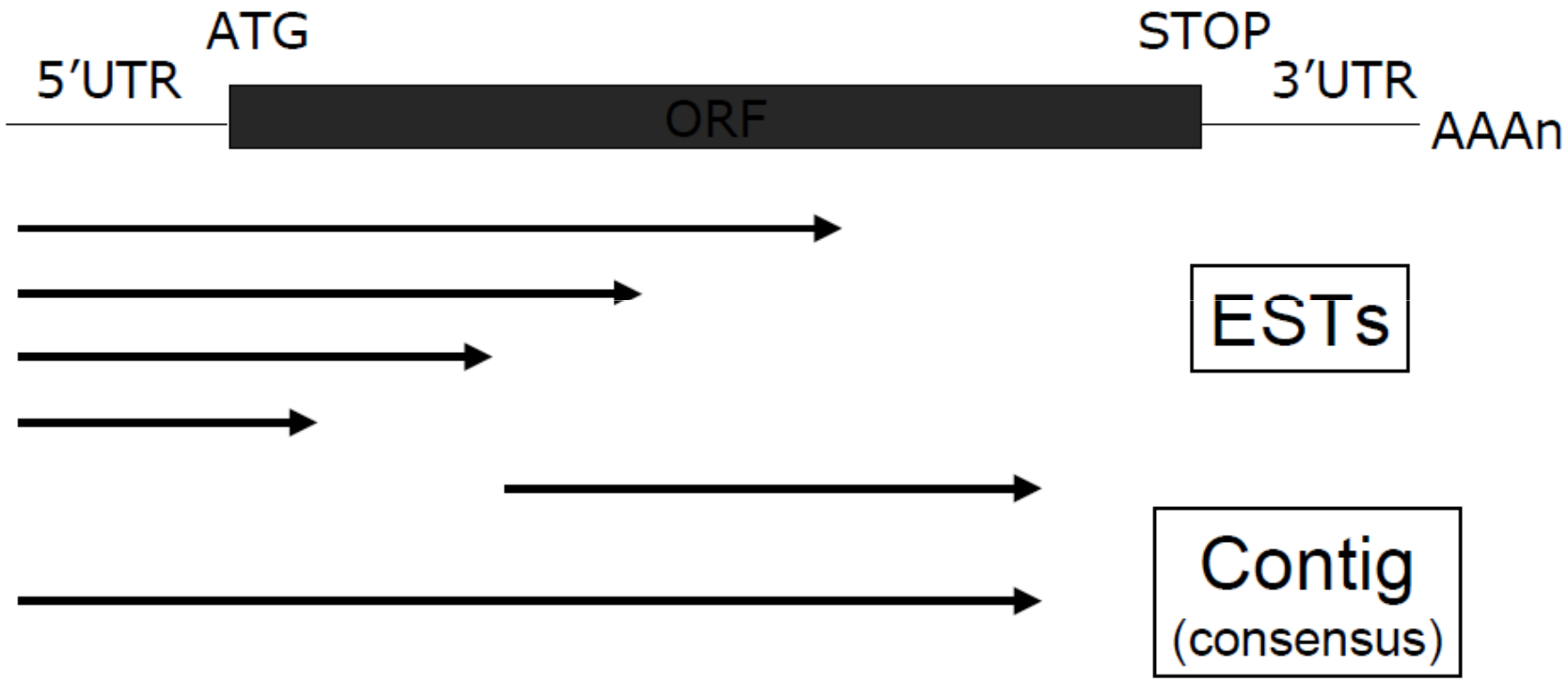
Nearly 30000 Arabidopsis thaliana EST (Expressed Sequence Tags) have been produced by a French and an American co Despite redundancy these sequences tag about half the expected Arabidopsis genes. Approximately 40% of the non-redu can be assigned a putative function by simple homology search. This programme allowed the identification of a large num which would have been very difficult to isolate by other classical techniques. It considerably stimulated many areas of pla the rapid discovery a large number of genes, by revealing multigene families and by allowing the analysis of differential ex the different members. Finally this programme facilitated construction of physical maps of the chromosomes and opened complete sequencing of the Arabidopsis genome and comparative mapping of the major plant crops.

Further progress towards a catalogue of all *Arabidopsis* genes: analysis of a set of 5000 non-redundant ESTs

Richard Cooke[1,], Monique Raynal[1], Michele Laudié[1], Françoise Grellet[1], Michel Delseny[1], Peter-Christian Morris[2], Danièle Guerrier[2], Jérôme Giraudat[2], Françoise Quigley[3], Gérard Clabault[3], You-Fang Li[3], Régis Mache[3], Micheline Krivitzky[4], Isabelle Jean-Jacques Gy[4], Martin Kreis[4], Alain Lecharny[4], Yves Parmentier[5], Jacqueline Marbach[5], Jacqueline Fleck[5], Bernadette Clément[6], Gabriel Philipps[6], Christine Hervé[7], Claude Bardet[7], Dominique Tremousaygue[7], Bernard Lescure[7], Christophe Lacomme[8], Dominique Roby[8], Marie-Françoise Jourjon[9], Patrick Chabrier[9], Jean-Louis Charpenteau[9], Thierry Desprez[10], Joelle Amselem[10], Helen Chiapello[10] and Herman Höfte[10]*

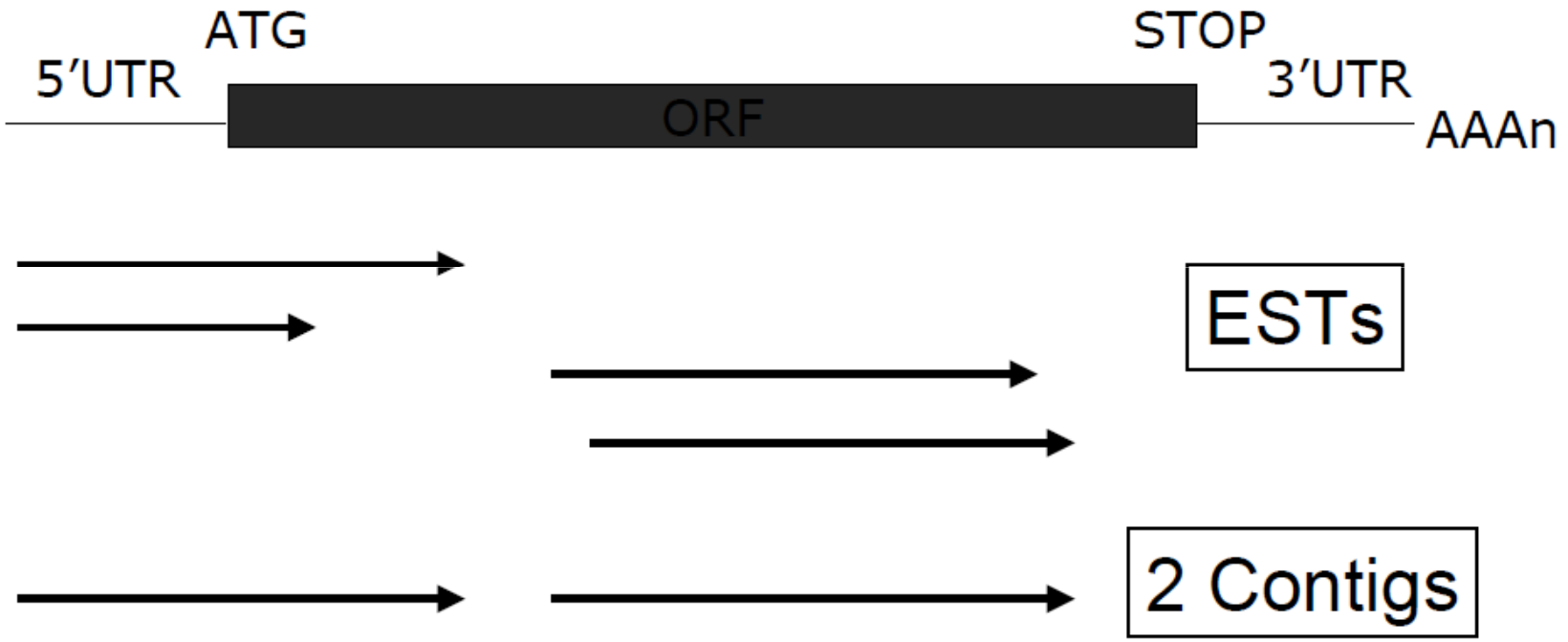
Nearly 7000 *Arabidopsis thaliana*-expressed sequence tags (ESTs) from 10 cDNA libraries have been sequenced, of which almost 5000 non-redundant tags have been submitted to the EMBL data bank. The quality of the cDNA libraries used is analysed. Similarity searches in international protein data banks have allowed the detection of significant similarities to a wide range of proteins from many organisms. Alignment with ESTs from the rice systematic sequencing project has allowed the detection of amino acid motifs which are conserved between the two organisms, thus identifying tags to genes encoding highly conserved proteins. These genes are candidates for a common framework in genome mapping projects in different plants.

Example: EST contig



Donc de plusieurs ESTs → 1 contig

Example: nonoverlapping ESTs



Mais aussi … plusieurs ESTs → … plusieurs contigs

De la sequence de l'Est à son annotation et à la base de données

Functional categories: example

01 Metabolism

01.01 Amino acid metabolism

01.02 Nitrogen and sulfur metab

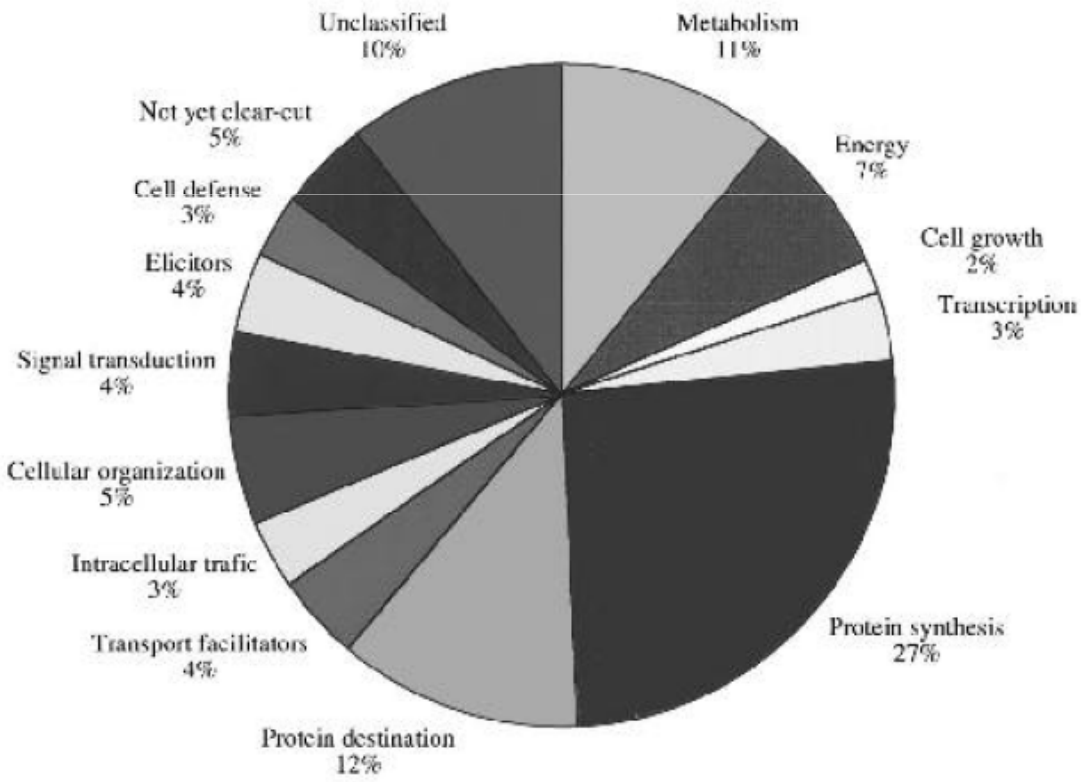
01.03 Nucleotide metabolism

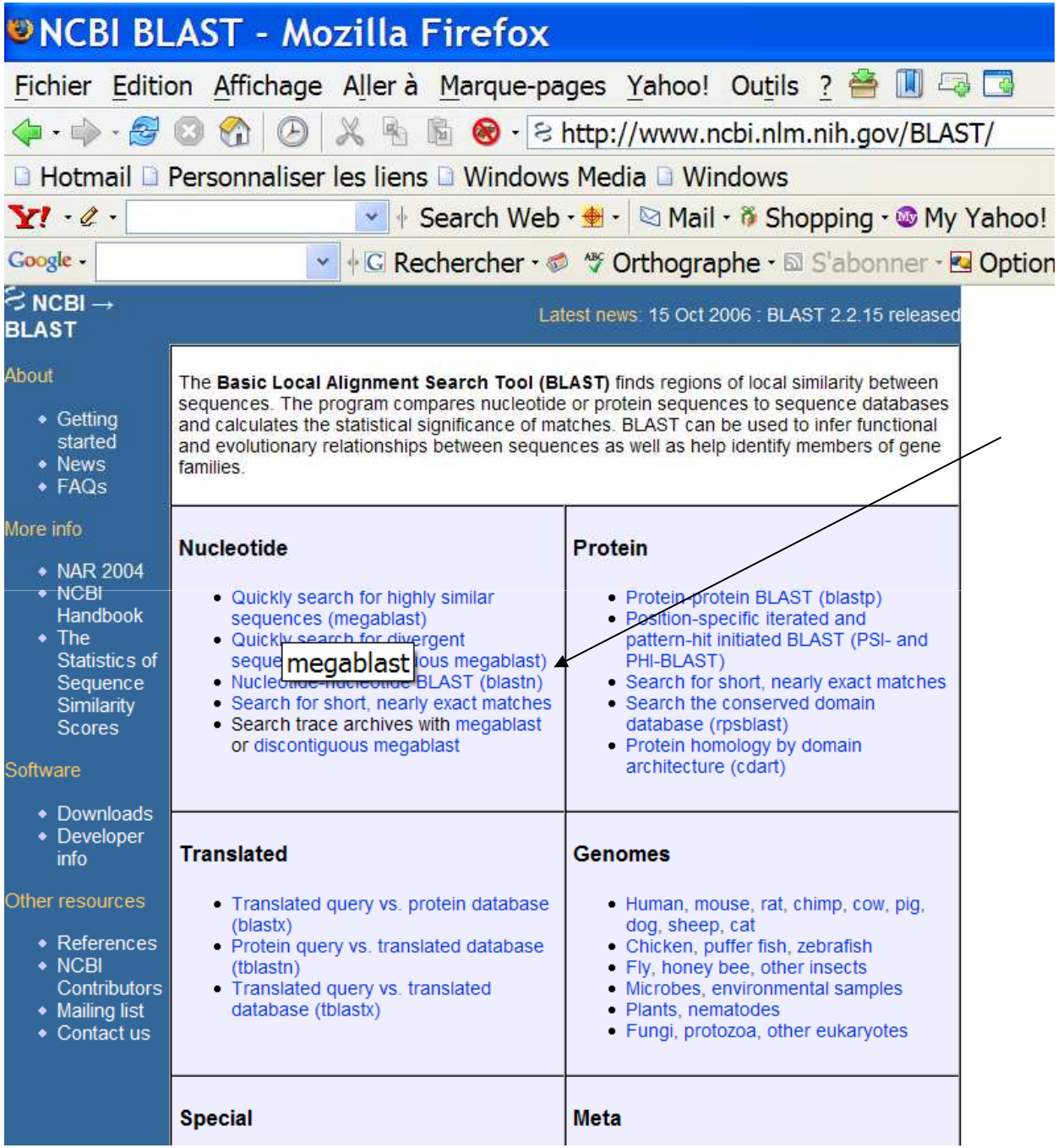
01.04 Phosphate metabolism

01.05 Carbohydrate metabolism

01.06 Lipid and sterol metabolis

01.07 Biosynthesis of vitamins, (
and prosthetic group



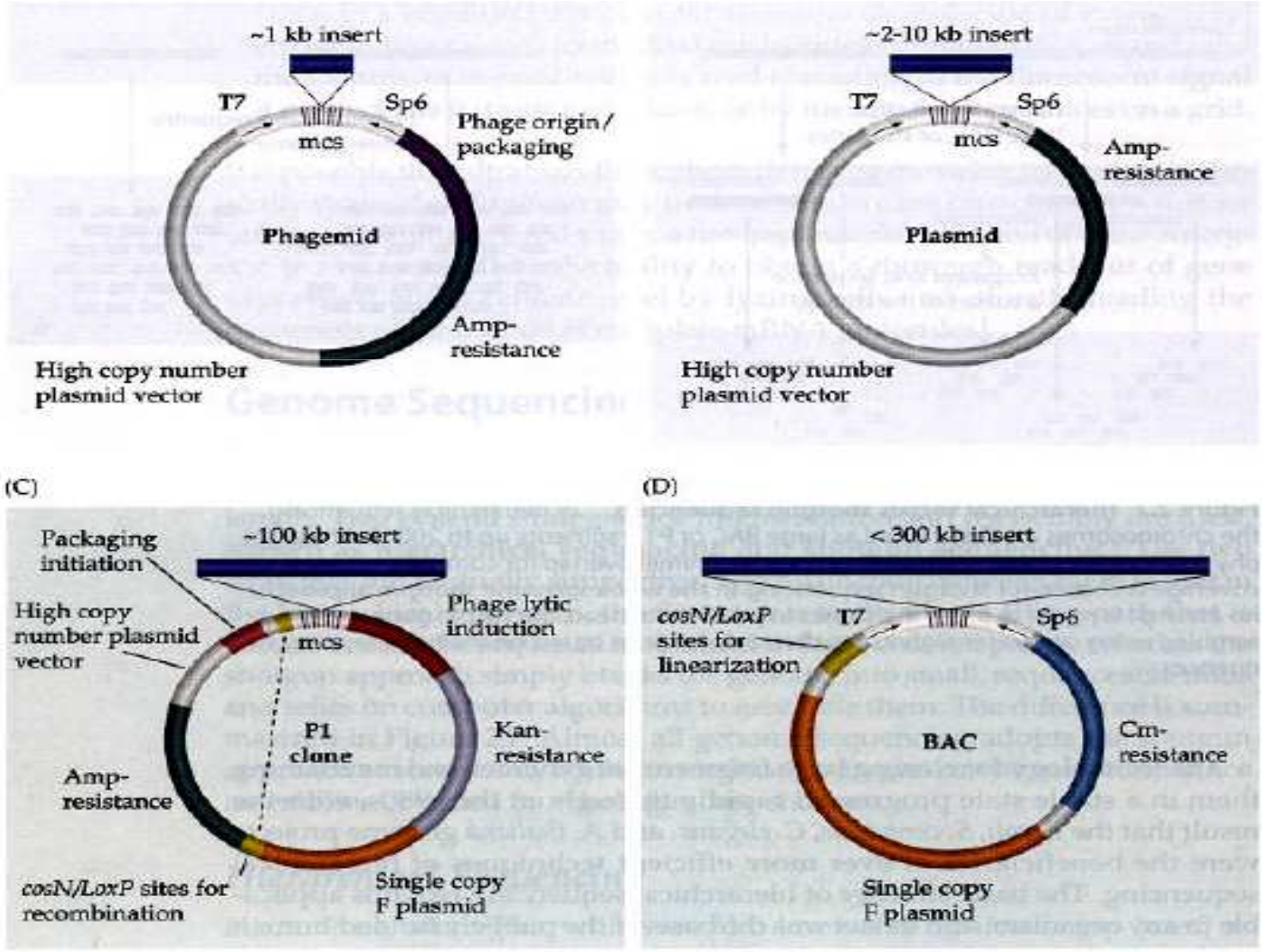

NCBI BLAST - Mozilla Firefox
Fichier Edition Affichage Aller à Marque-pages Yahoo! Outils ?
http://www.ncbi.nlm.nih.gov/BLAST/
Hotmail Personnaliser les liens Windows Media Windows
Search Web Mail Shopping My Yahoo!
Google Rechercher Orthographe S'abonner Option
NCBI → BLAST
Latest news: 15 Oct 2006 : BLAST 2.2.15 released
About
Getting started
News
FAQs
More info
NAR 2004
NCBI Handbook
The Statistics of Sequence Similarity Scores
Software
Downloads
Developer info
Other resources
References
NCBI Contributors
Mailing list
Contact us
The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.
Nucleotide
Quickly search for highly similar sequences (megablast)
Quickly search for divergent
megablast
Nucleotide-nucleotide BLAST (blastn)
Search for short, nearly exact matches
Search trace archives with megablast or discontiguous megablast
Protein
Protein-protein BLAST (blastp)
Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)
Search for short, nearly exact matches
Search the conserved domain database (rpsblast)
Protein homology by domain architecture (cdart)
Translated
Translated query vs. protein database (blastx)
Protein query vs. translated database (tblastn)
Translated query vs. translated database (tblastx)
Genomes
Human, mouse, rat, chimp, cow, pig, dog, sheep, cat
Chicken, puffer fish, zebrafish
Fly, honey bee, other insects
Microbes, environmental samples
Plants, nematodes
Fungi, protozoa, other eukaryotes
Special
Meta

Le séquençage des génomes

Major differences between prokaryote and eukaryote genome

	Genome size	Chromosome	Centromere	Telemere	Organization	Repetitive Sequence
Prokaryotes	Small	Single circular, few linear	No	No	High gene density &lack introns	Non or very low
Eukaryotes	Large	Linear	yes	yes	Low gene density & disrupted by introns	High

- Le développement de vecteurs pour le séquençage des génomes



Bacterial Artificial Chromosome (BAC)

- **Bacterial cloning system based on E. coli single copy F factor (PNAS 1992. 89: 8794-8797)**
- **Easy to manipulate**
- **Easy to screen**
- **Stable maintenance of cloned DNA**
- **Nonchimeric**
- **High transformation efficiency**

BAC : 100à 350kb

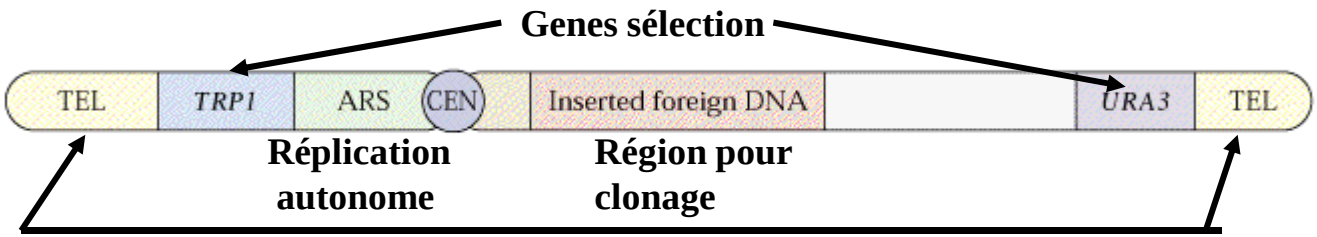
Repose sur episome F, présent en copie unique chez certaines souches d'E Coli.

→pb de conservation or pas de recombinaison et monocopie

YAC : de 250kb à 2900 kb !

Propagation possible d'ADN exogene dans la levure sous forme d'un chromosome artificiel

→faible efficacite de transfo, pb de recombinaison, clonage de grands fragments



Génome humain :

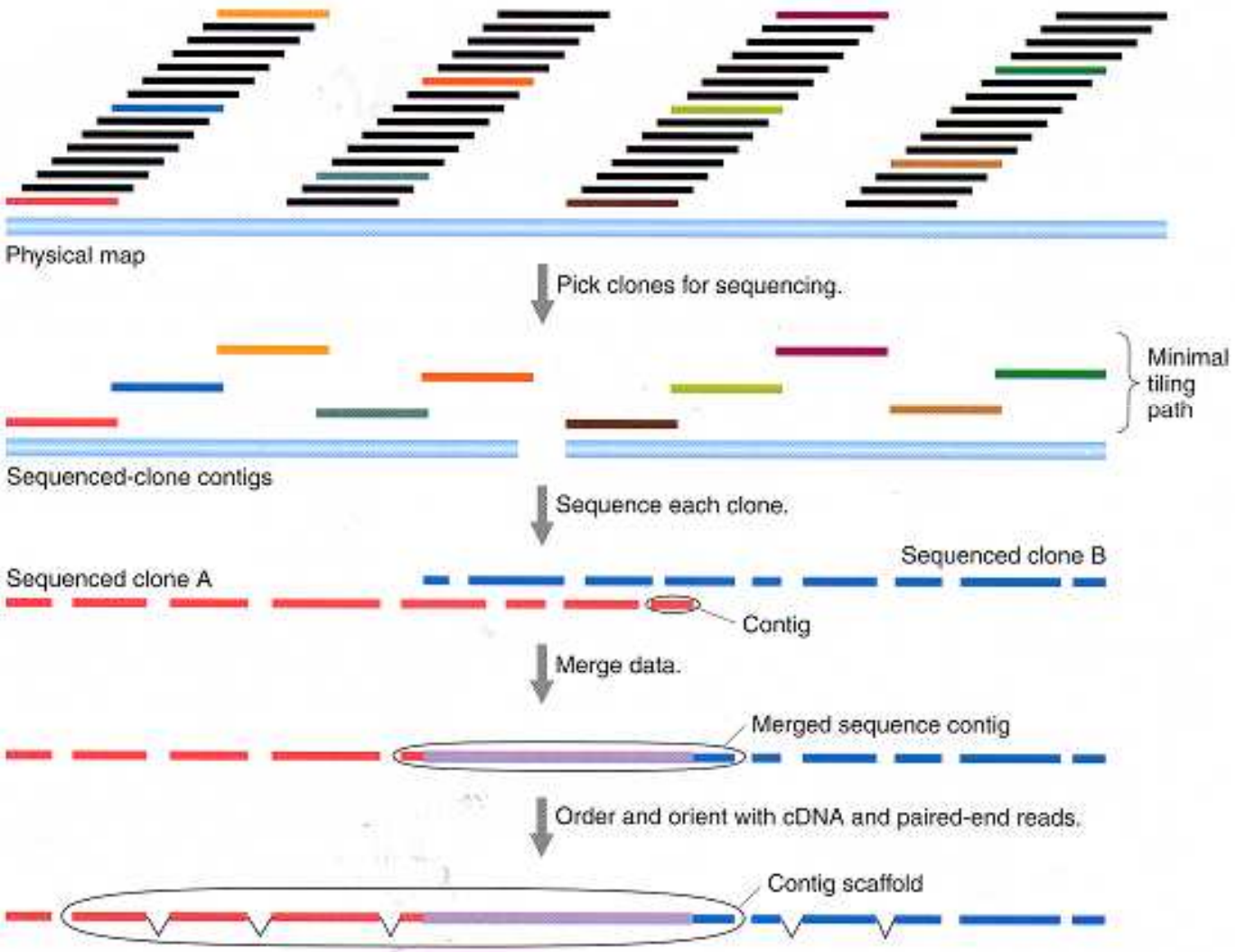
3,2 milliards de nucléotides, soit, en caractères, le contenu de 2000 livres de 500 p.

Pour couvrir une banque génomique humaine :

920 000 clones	phages lambda 15kb
307 000	cosmides 45 kb
92 000	BAC 150kb
14 000	YAC 1000kb

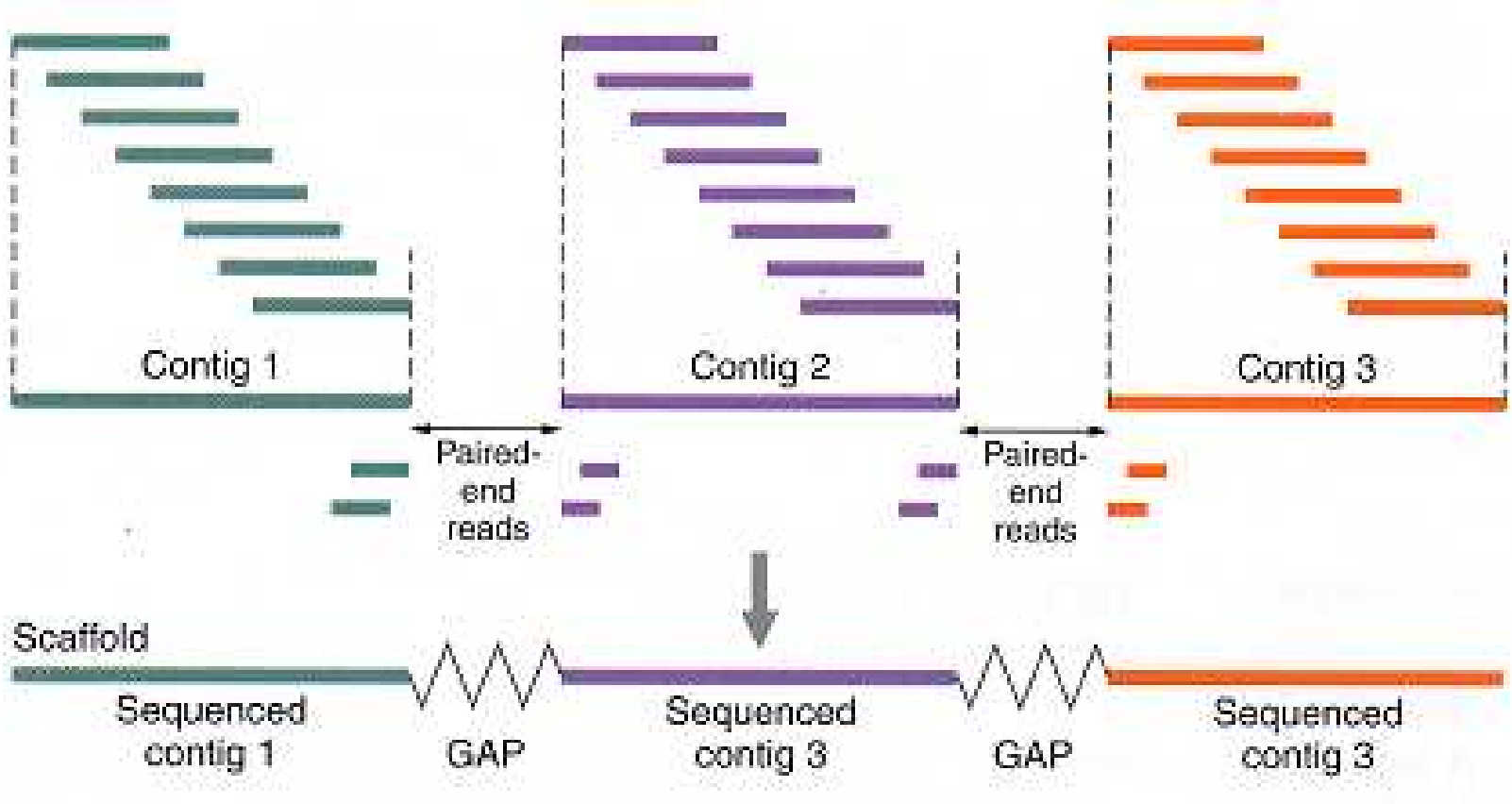
Genome Sequencing by 1st Physical Mapping (Clone à clone)

Extraction A DN, digestion, clonage dans vecteur adapté, assemblage des clones



Genome Sequencing by “Shotgun” Method

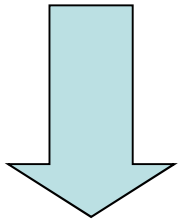
Extraction ADN, digestion en fragment de 1 a 10 kb, sequencage (6-8 fois genome), assemblage



<u>Méthodes de séquençage reposent sur la technique de SANGER</u>

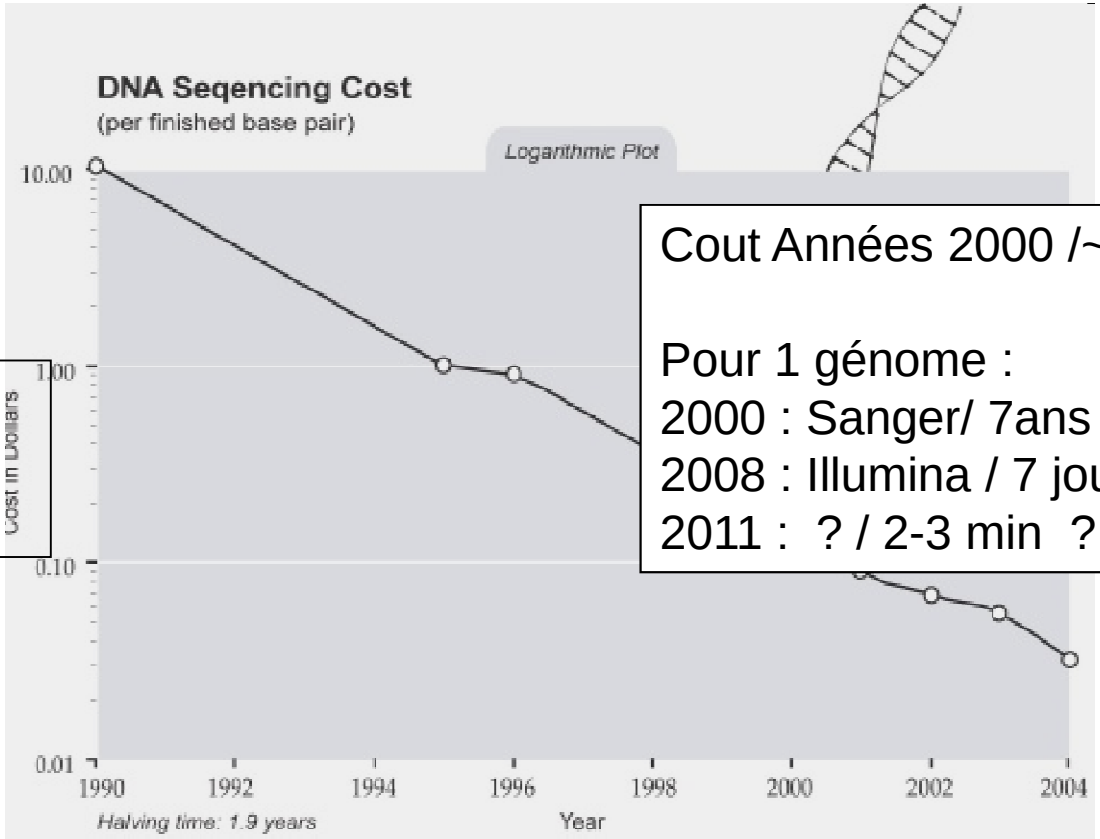
Points -
- **Nécessite d'isoler un clone par PCR, clonage … avant de séquencer**
- **time consuming → donc cher**

Points +
- **mais génère des fragments de 700 à 1000 bp avec peu d'erreurs**



UHTS

ULTRA-HIGH-THROUGHPUT SEQUENCING



Cout Années 2000 /~ 1000 depuis 2008 …

Pour 1 génome :
2000 : Sanger/ 7ans / 500 personnes/ 70 000 000 dollars
2008 : Illumina / 7 jours /2 post –docs / 7000 dollars
2011 : ? / 2-3 min ? / 70 dollars

Cout

Taille

2009

Table 1 Second-generation DNA sequencing technologies

	Feature generation	Sequencing by synthesis	Cost per megabase	Cost per instrument	Paired ends?	1° error modality	Read-length	References
454	Emulsion PCR	Polymerase (pyrosequencing)	~$60	$500,000	Yes	Indel	250 bp	14,20
Solexa	Bridge PCR	Polymerase (reversible terminators)	~$2	$430,000	Yes	Subst.	36 bp	17,22
SOLiD	Emulsion PCR	Ligase (octamers with two-base encoding)	~$2	$591,000	Yes	Subst.	35 bp	13,26
Polonator	Emulsion PCR	Ligase (nonamers)	~$1	$155,000	Yes	Subst.	13 bp	13,20
HeliScope	Single molecule	Polymerase (asynchronous extensions)	~$1	$1,350,000	Yes	Del	30 bp	18,30

The pace with which the field is moving makes it likely that estimates for costs and read-lengths will be quickly outdated. Vendors including Roche Applied Science, Illumina, and Applied Biosystems have major upgrade releases currently in progress. Estimated costs-per-megabase are approximate and inclusive only of reagents. Read-lengths are for single tags. Subst., substitutions; indel, insertions or deletions; del, deletions.

Pour quelles applications ?

Table 2 Applications of next-generation sequencing

Category	Examples of applications	Refs
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes	44
Reduced representation sequencing	Large-scale polymorphism discovery	45
Targeted genomic resequencing	Targeted polymorphism and mutation discovery	46–52
Paired end sequencing	Discovery of inherited and acquired structural variation	53,54
Metagenomic sequencing	Discovery of infectious and commensal flora	55
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations	56–63
Small RNA sequencing	microRNA profiling	64
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA	60,65,66
Chromatin immunoprecipitation–sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions	67–70
Nuclease fragmentation and sequencing	Nucleosome positioning	69
Molecular barcoding	Multiplex sequencing of samples from multiple individuals	61,71

La bonne question ?

iotechnology

What would you do if you could sequence everything?

Avak Kahvejian[1], John Quackenbush[2] & John F Thompson[1]

Nature biotech, 2008

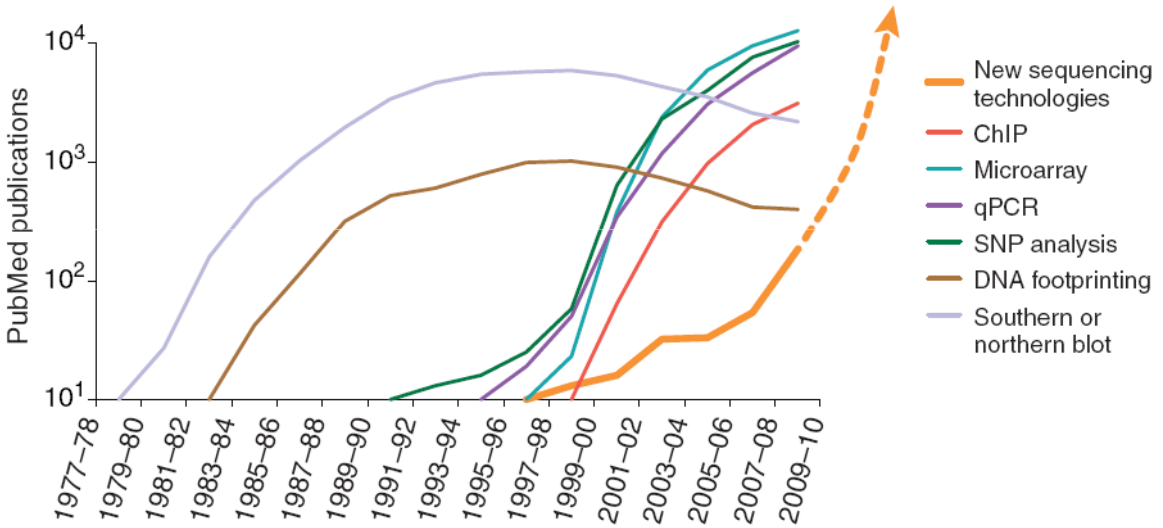


Figure 1 The number of publications with keywords for nucleic acid detection and sequencing technologies. PubMed (http://www.ncbi.nlm.nih.gov/sites/entrez) was searched in two-year increments for key words and the number of hits plotted over time. For 2007–2008, results from January 1–March 31, 2008 were multiplied by four and added to those for 2007. Key words used were those listed in the legend except for new sequencing technologies ('next-generation sequencing' or 'high-throughput sequencing'), ChIP ('chromatin immunoprecipitation' or 'ChIP-Chip' or 'ChIP-PCR' or 'ChIP-Seq'), qPCR (TaqMan or qPCR or 'real-time PCR') and SNP analysis (SNPs or 'single-nucleotide polymorphisms' and not nitroprusside (nitroprusside is excluded because sodium nitroprusside is sometimes abbreviated as 'SNP' but is generally unrelated to genetics)).

ire.com/naturebiotechnology

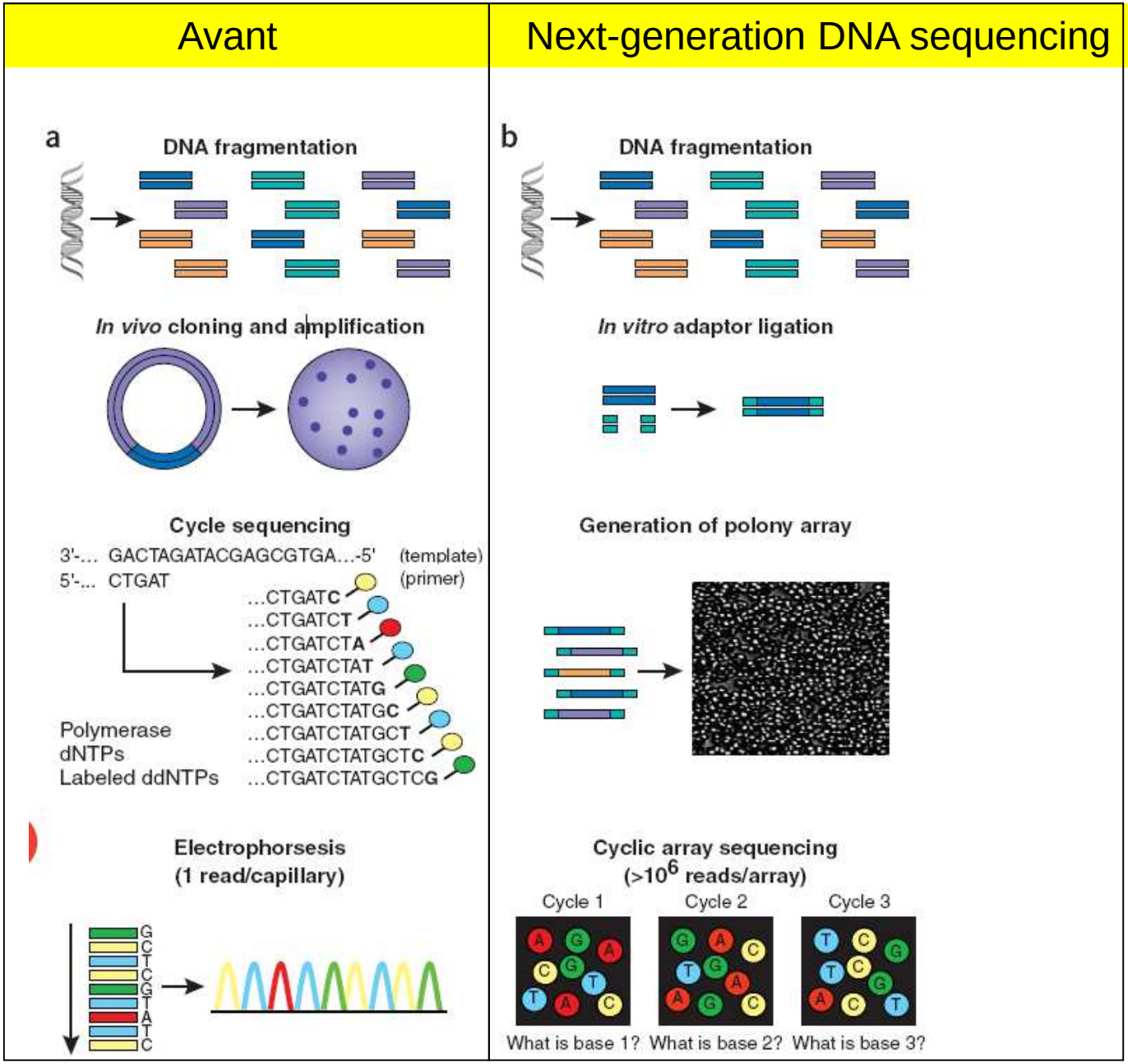
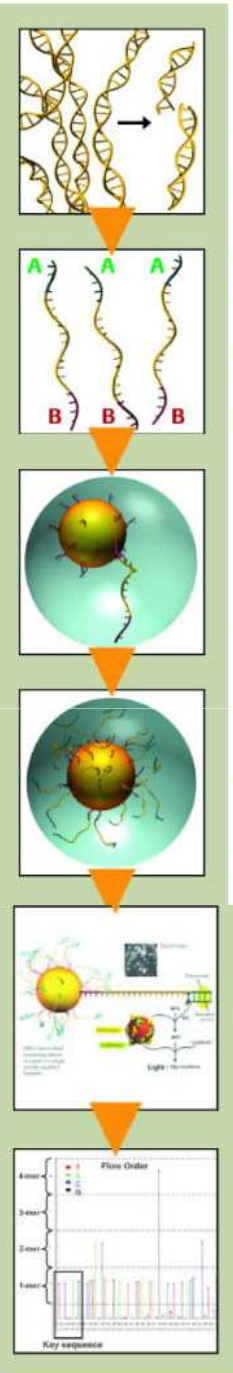


Figure 1 Work flow of conventional versus second-generation sequencing. (a) With high-throughput

454 pyrosequencing



Sample Input and Fragmentation

The Genome Sequencer FLX System supports the sequencing of samples from a wide variety of starting materials including genomic DNA, PCR products, BACs, and cDNA. Samples such as genomic DNA and BACs are fractionated into small, 300- to 800-basepair fragments. For smaller samples, such as small non-coding RNA or PCR amplicons, fragmentation is not required. Instead, short PCR products amplified using Genome Sequencer fusion primers can be used for immobilization onto DNA capture beads as shown below under "One Fragment = One Bead".

The GS FLX System supports multiple sample prep options, click here to see the full list.

Library Preparation

Using a series of standard molecular biology techniques, short adaptors (A and B) - specific for both the 3' and 5' ends - are added to each fragment. The adaptors are used for purification, amplification, and sequencing steps. Single-stranded fragments with A and B adaptors compose the sample library used for subsequent workflow steps.

One Fragment = One Bead

The single-stranded DNA library is immobilized onto specifically designed DNA Capture Beads. Each bead carries a unique single-stranded DNA library fragment. The bead-bound library is emulsified with amplification reagents in a water-in-oil mixture resulting in microreactors containing just one bead with one unique sample-library fragment.

emPCR (Emulsion PCR) Amplification

Each unique sample library fragment is amplified within its own microreactor, excluding competing or contaminating sequences. Amplification of the entire fragment collection is done in parallel; for each fragment, this results in a copy number of several million per bead. Subsequently, the emulsion PCR is broken while the amplified fragments remain bound to their specific beads.

One Bead = One Read

The clonally amplified fragments are enriched and loaded onto a PicoTiterPlate device for sequencing. The diameter of the PicoTiterPlate wells allows for only one bead per well. After addition of sequencing enzymes, the fluidics subsystem of the Genome Sequencer FLX Instrument flows individual nucleotides in a fixed order across the hundreds of thousands of wells containing one bead each. Addition of one (or more) nucleotide(s) complementary to the template strand results in a chemiluminescent signal recorded by the CCD camera of the Genome Sequencer FLX Instrument. For a detailed explanation of this reaction see Sequencing Chemistry.

Data Analysis

The combination of signal intensity and positional information generated across the PicoTiterPlate device allows the software to determine the sequence of more than 1,000,000 individual reads per 10-hour instrument run simultaneously. For sequencing-data analysis, three different bioinformatics tools are available supporting the following applications: de novo assembly up to 400 megabases; resequencing genomes of any size; and amplicon variant detection by comparison with a known reference sequence.

Fragmentation ADN

Ajout d'adaptateur

Couplage a des billes
(1 fragment = 1 bille)
Puis billes dans µplaque

PCR

Pyrosequencage : A chaque cycle on ajoute 1 dNTP (dATP, ou dGTP, ou dCTP, ou dTTP)

Si il est incorporé alors pyrophosphate est libéré et 1 signal est enregistré grace à un systeme luminescent

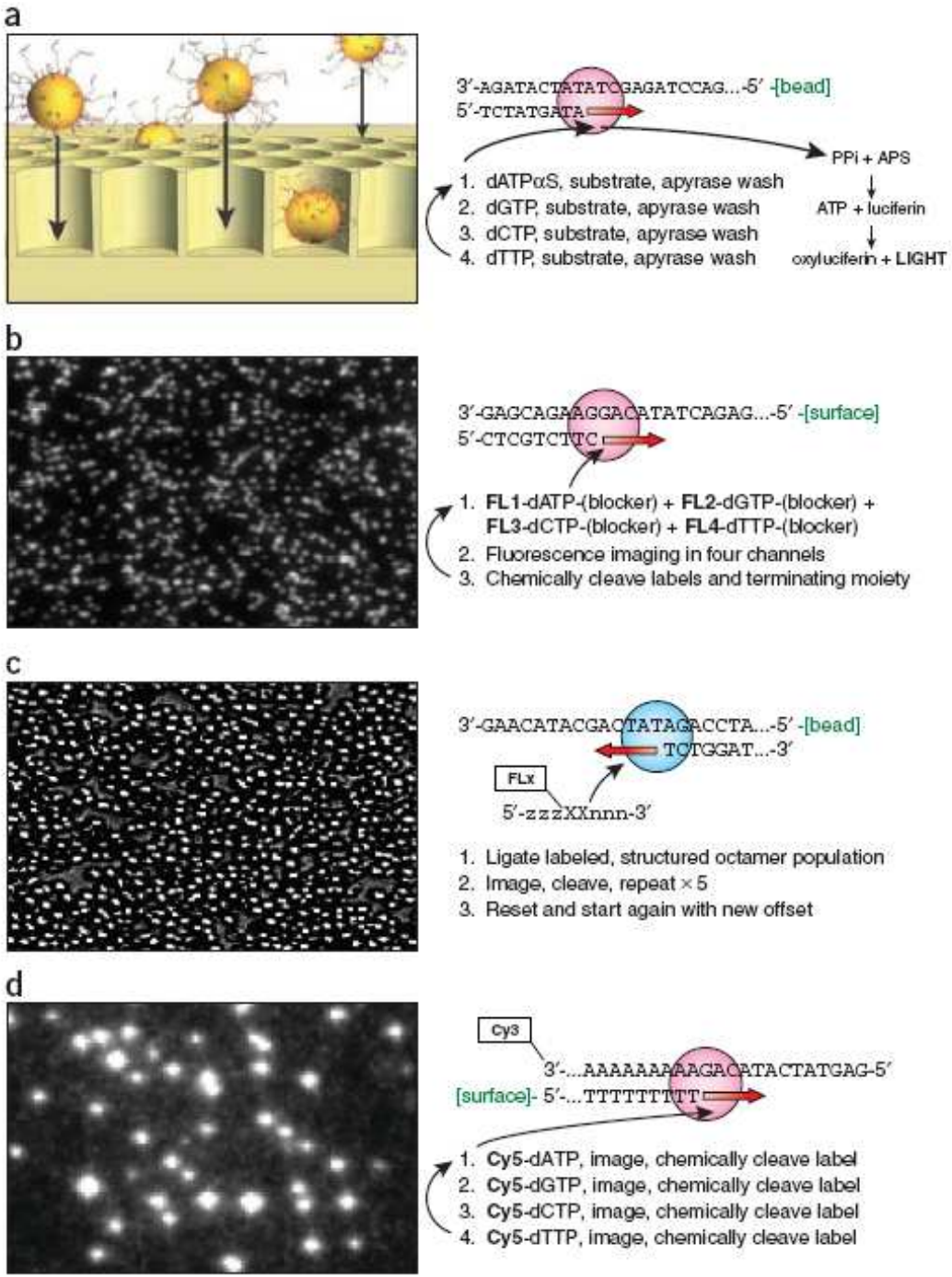
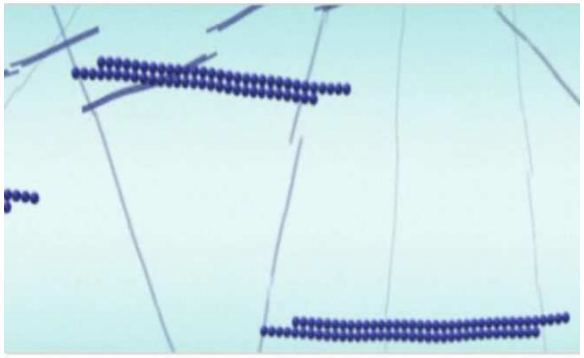


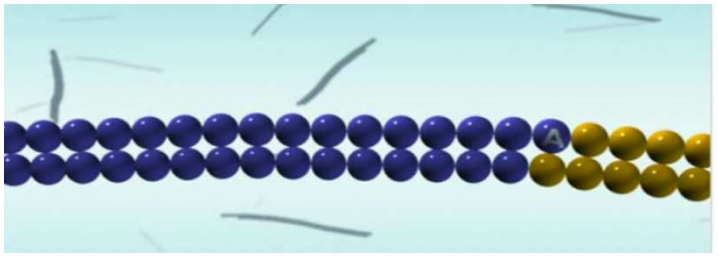
Illustration 454

ILLUMINA

1. DNA fragmentation →

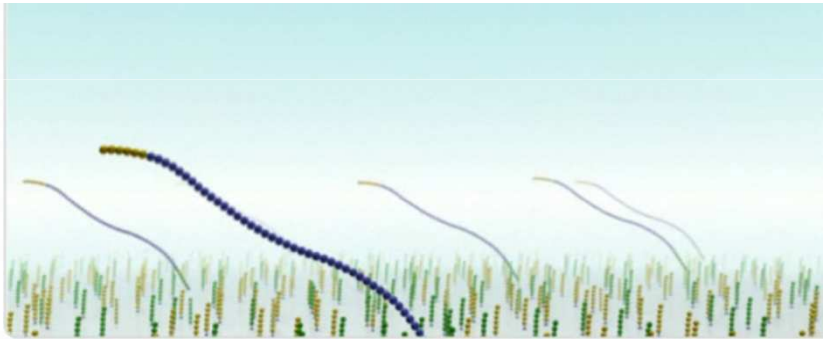


2. ajout d'adaptateur →

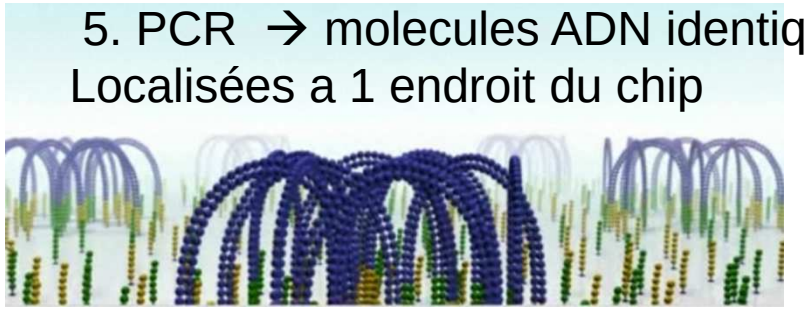


3. Chips avec oligos complementaires → aux adaptateurs

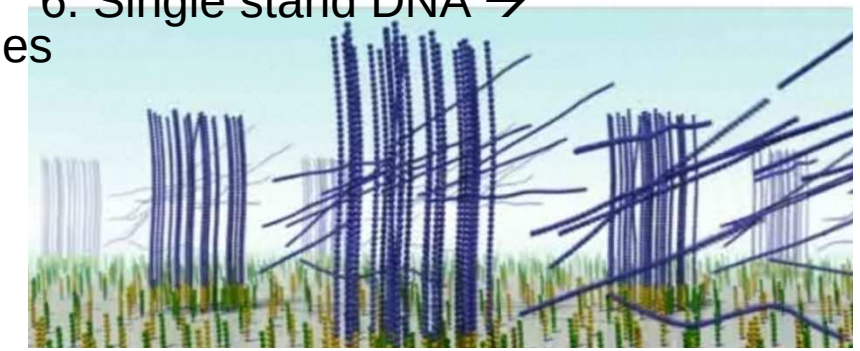
4. Fixation ADN au chip



5. PCR → molecules ADN identiques Localisées a 1 endroit du chip



6. Single stand DNA →



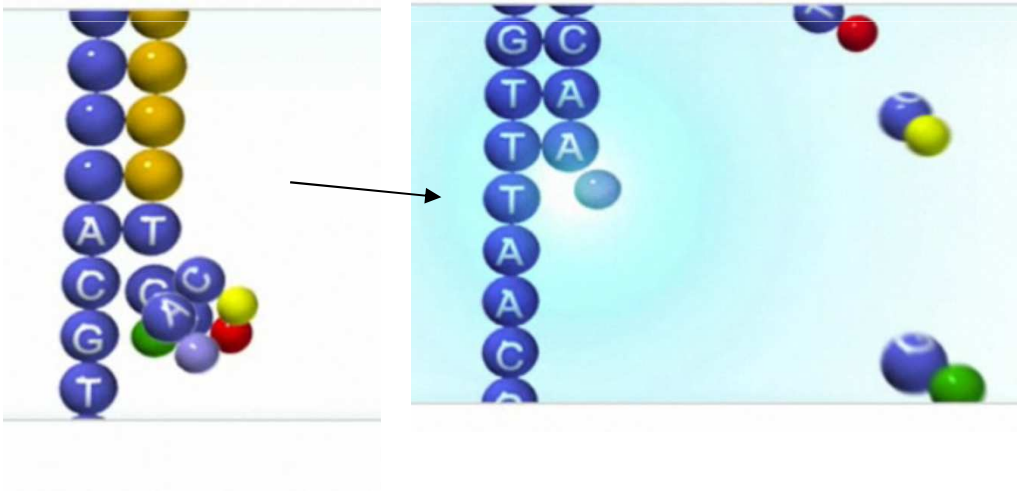
ILLUMINA

Sequencage

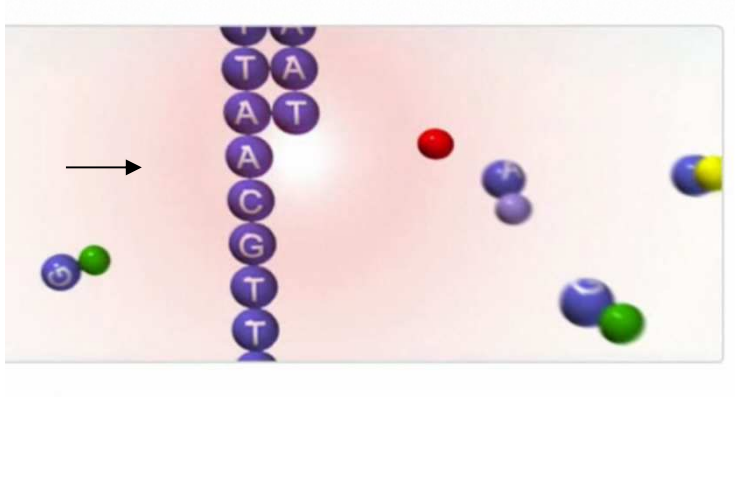
7. Ajout des 4 dNTP (chacun avec un fluorochrome spécifique)
Séquencage en Synthèse

8. Fixation du dNTP libération du fluorochrome

9.Détection

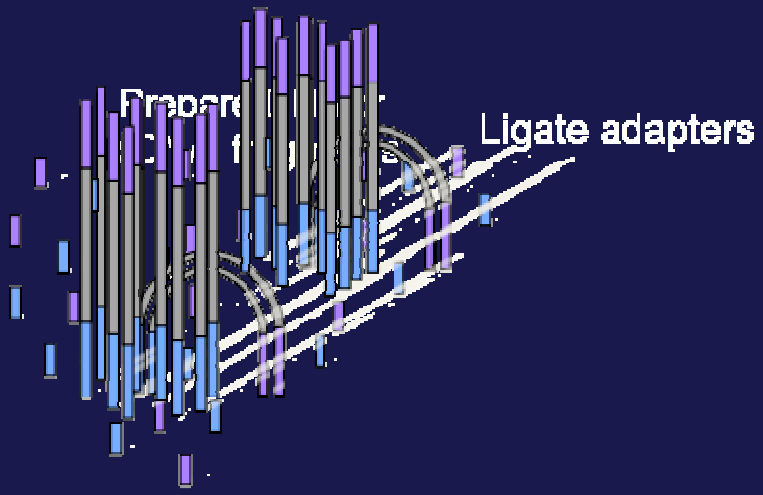


A…



AT… etc …

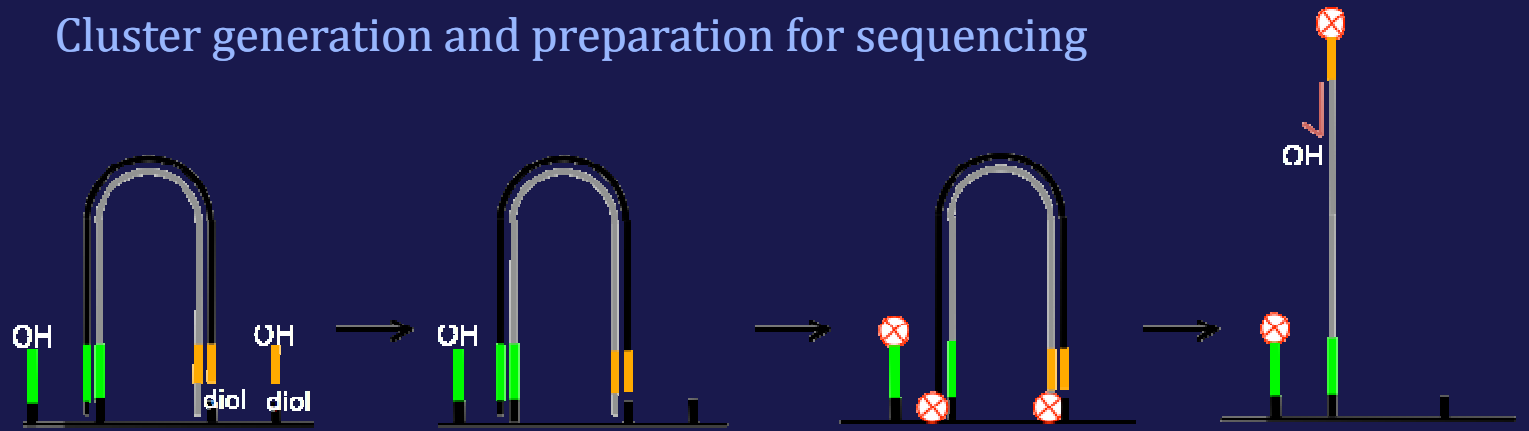
Sample generation and cluster generation



Attach single molecules to surface

Amplify to form clusters

Cluster generation and preparation for sequencing



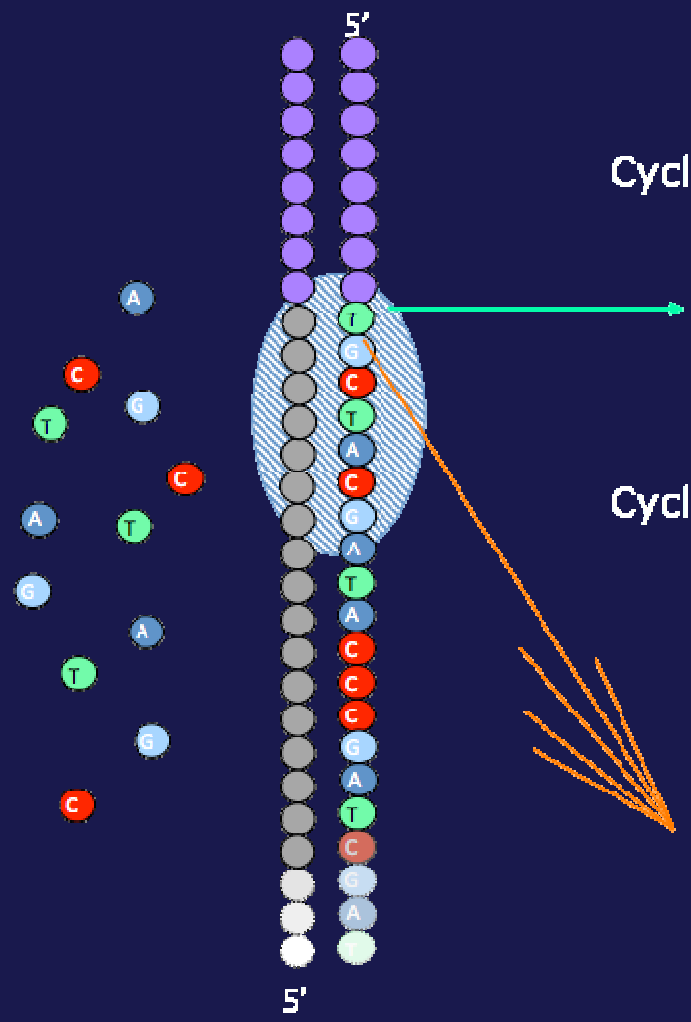
Cluster Amplification

Periodate Linearization

Blocking with ddNTP (⊗)

Denature and Hybridization of Sequencing primer

Sequencing by Synthesis (SBS)



Cycle 1: Add sequencing reagents

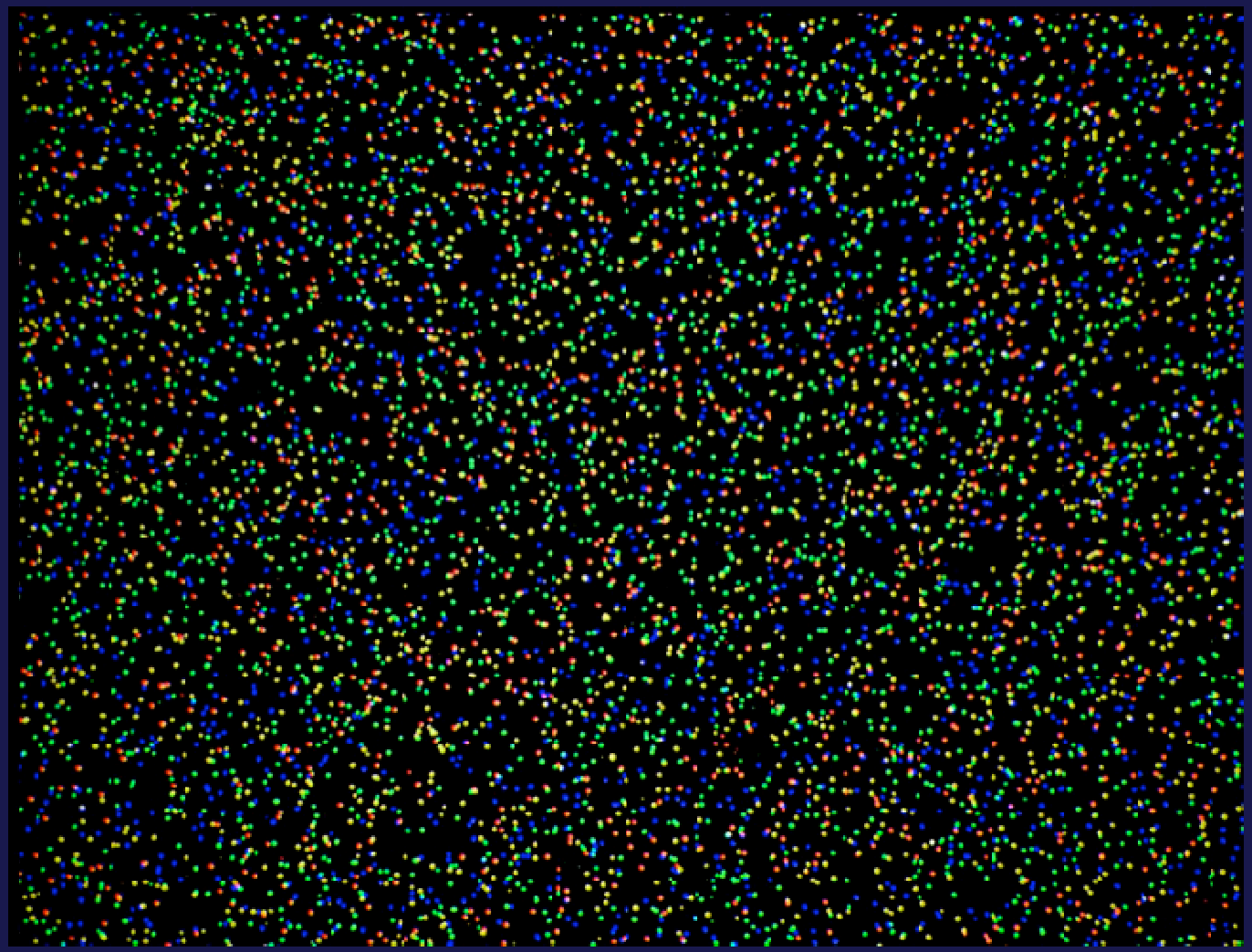
First base incorporated

Detect Signal

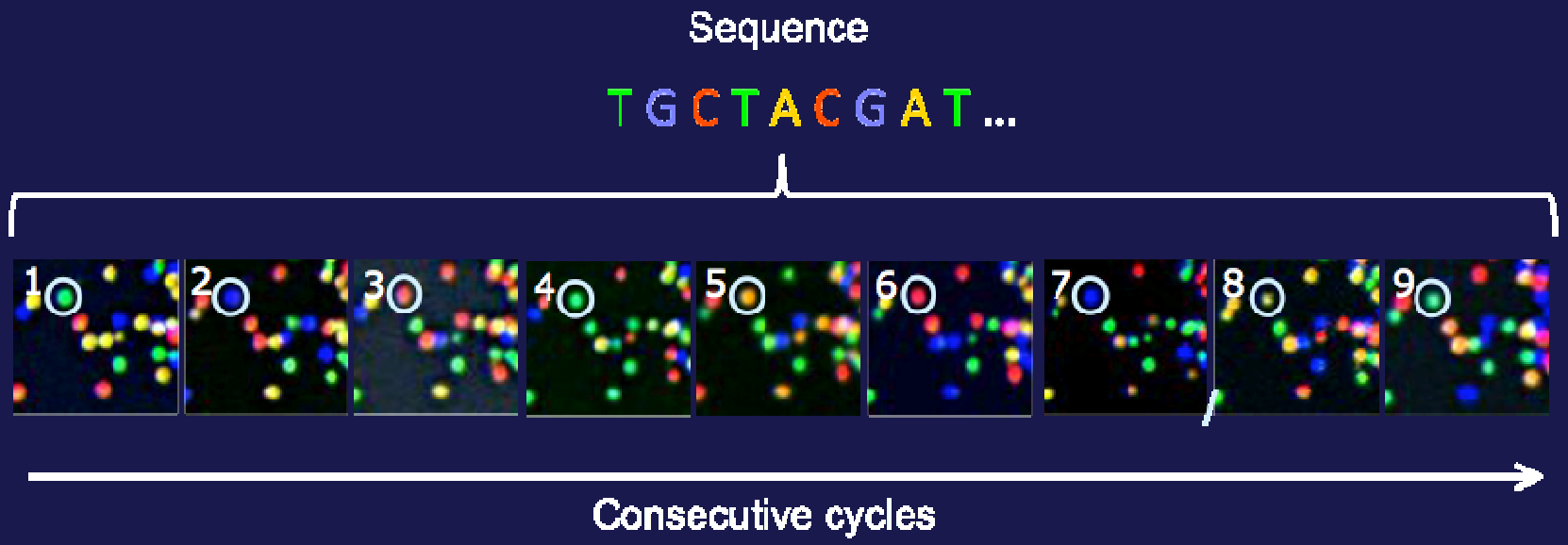
Cleave Terminator and Dye

Cycle 2-n: Add sequencing reagents and repeat

Sequencing by Synthesis (SBS)



Base Calling



The identity of each base of a cluster is read from stacked sequential images

doi:10.1038/nature10242

An integrated semiconductor device enabling non-optical genome sequencing

Jonathan M. Rothberg[1], Wolfgang Hinz[1], Todd M. Rearick[1], Jonathan Schultz[1], William Mileski[1], Mel Davey[1], John H. Leamon[1], Kim Johnson[1], Mark J. Milgrew[1], Matthew Edwards[1], Jeremy Hoon[1], Jan F. Simons[1], David Marran[1], Jason W. Myers[1], John F. Davidson[1], Annika Branting[1], John R. Nobile[1], Bernard P. Puc[1], David Light[1], Travis A. Clark[1], Martin Huber[1], Jeffrey T. Branciforte[1], Isaac B. Stoner[1], Simon E. Cawley[1], Michael Lyons[1], Yutao Fu[1], Nils Homer[1], Marina Sedova[1], Xin Miao[1], Brian Reed[1], Jeffrey Sabina[1], Erika Feierstein[1], Michelle Schorn[1], Mohammad Alanjary[1], Eileen Dimalanta[1], Devin Dressman[1], Rachel Kasinskas[1], Tanya Sokolsky[1], Jacqueline A. Fidanza[1], Eugeni Namsaraev[1], Kevin J. McKernan[1], Alan Williams[1], G. Thomas Roth[1] & James Bustillo[1]

Nature 2011, 475, 348-352

DNA sequencing technology in semiconductor able to directly perform non-optical DNA sequencing of genomes.

Sequence data are obtained by directly sensing the ions produced by template-directed DNA polymerase synthesis using all-natural nucleotides on this massively parallel semiconductor-sensing device or ion chip.

The ion chip contains ion-sensitive, field-effect transistor-based sensors in perfect register with 1.2 million wells,
which provide confinement and allow parallel, simultaneous detection of independent sequencing reactions.

We show the performance of the system by sequencing three bacterial genomes, its robustness and scalability
by producing ion chips with up to 10 times as many sensors and sequencing a human genome.

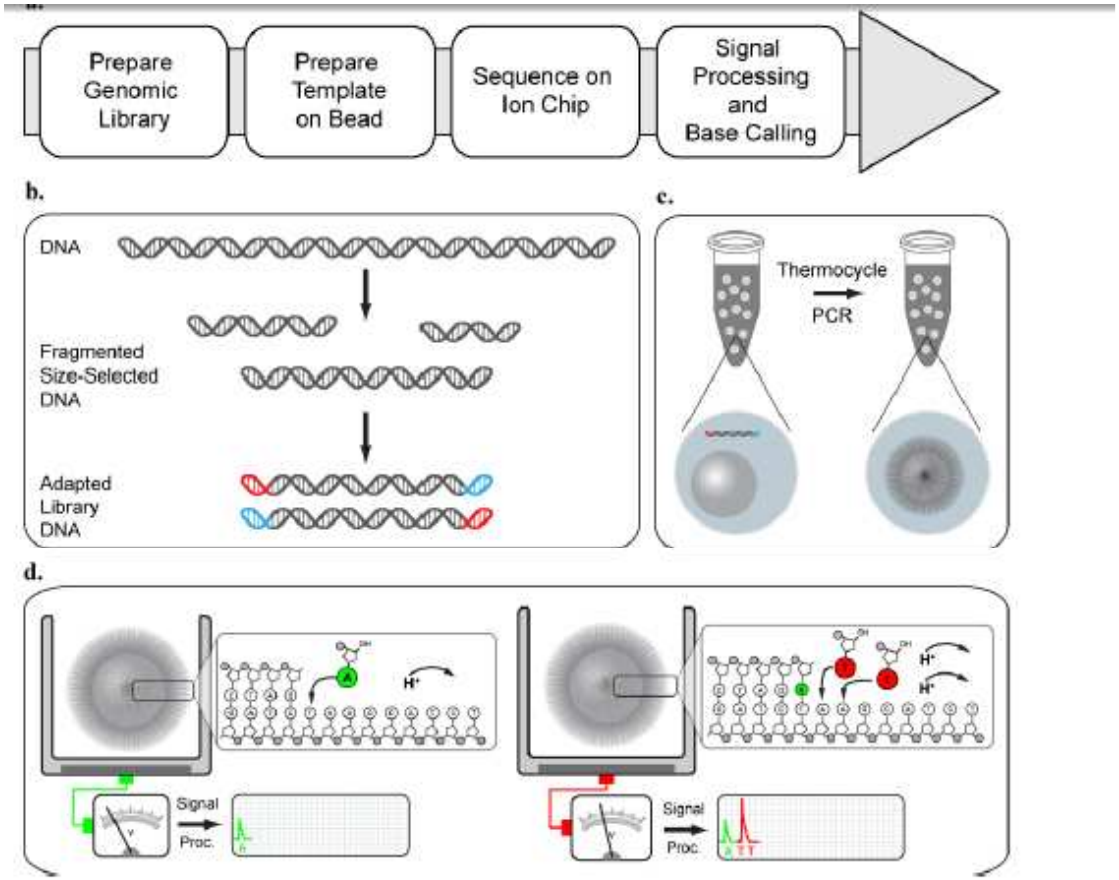
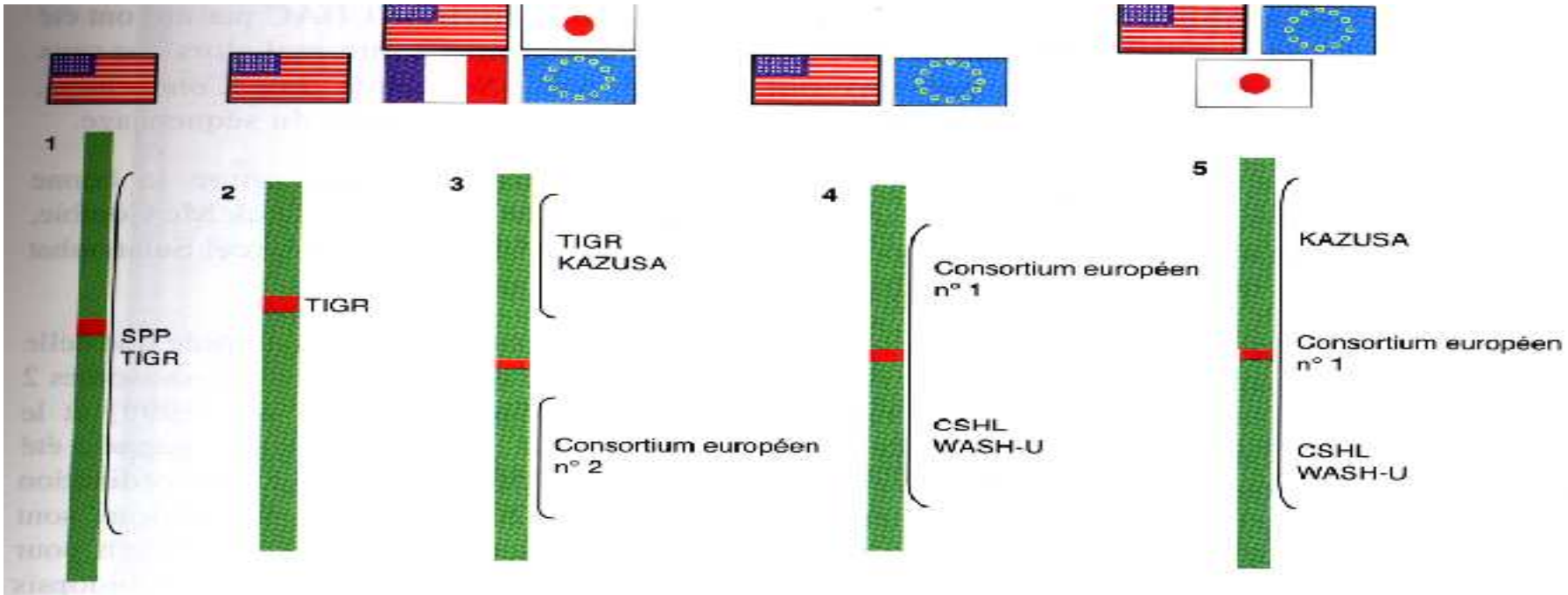


Figure S1 Process overview

a, Overview of ion sequencing work flow. **b,** Prepare genomic library, DNA is fragmented, sized, and forward and reverse adapters ligated. **c,** Amplify Template on bead, adapter-ligated libraries are clonally amplified onto beads. A magnetic bead-based enrichment process selects template-carrying beads. **d,** Sequence on ion chip, sequencing primers and DNA polymerase are bound to the template-carrying beads, beads are pipetted into the chip's loading port. The chip is installed in the sequencing instrument; all four nucleotides cyclically flowed in an automated 2-hour run. Signal processing, software converts the raw data into measurements of incorporation in each well for each successive nucleotide flow. After bases are called, each read is passed through a filter to exclude low-accuracy reads and per-base quality values are predicted.

Séquençage des génomes : Le Yalta des chromosomes chez Arabidopsis

Arabidopsis thaliana



SPP Consortium : Stanford DNA Sequencing & Technology Center, Stanford University, *Arabidopsis thaliana* Genome Center, University of Pennsylvania, Plant Gene Expression Center, USDA.
TIGR : The Institute for Genomic Research, Washington.
KAZUSA DNA Research Institute, Japan.
Consortium européen n°1 : coordinateur : John Innes Center, Norwich.
Consortium européen n°2 : coordinateur : Génoscope, Évry.
Wash-U : Washington University, St Louis.
CSHL : Cold Spring Harbor Laboratory, NY.

Figure 6. Répartition du séquençage des 5 chromosomes d'*Arabidopsis thaliana* entre les différents centres de séquençage de l'AGI.

Stratégie clone à clone

Séquençage des génomes : Le Yalta des chromosomes

Oriza sativa

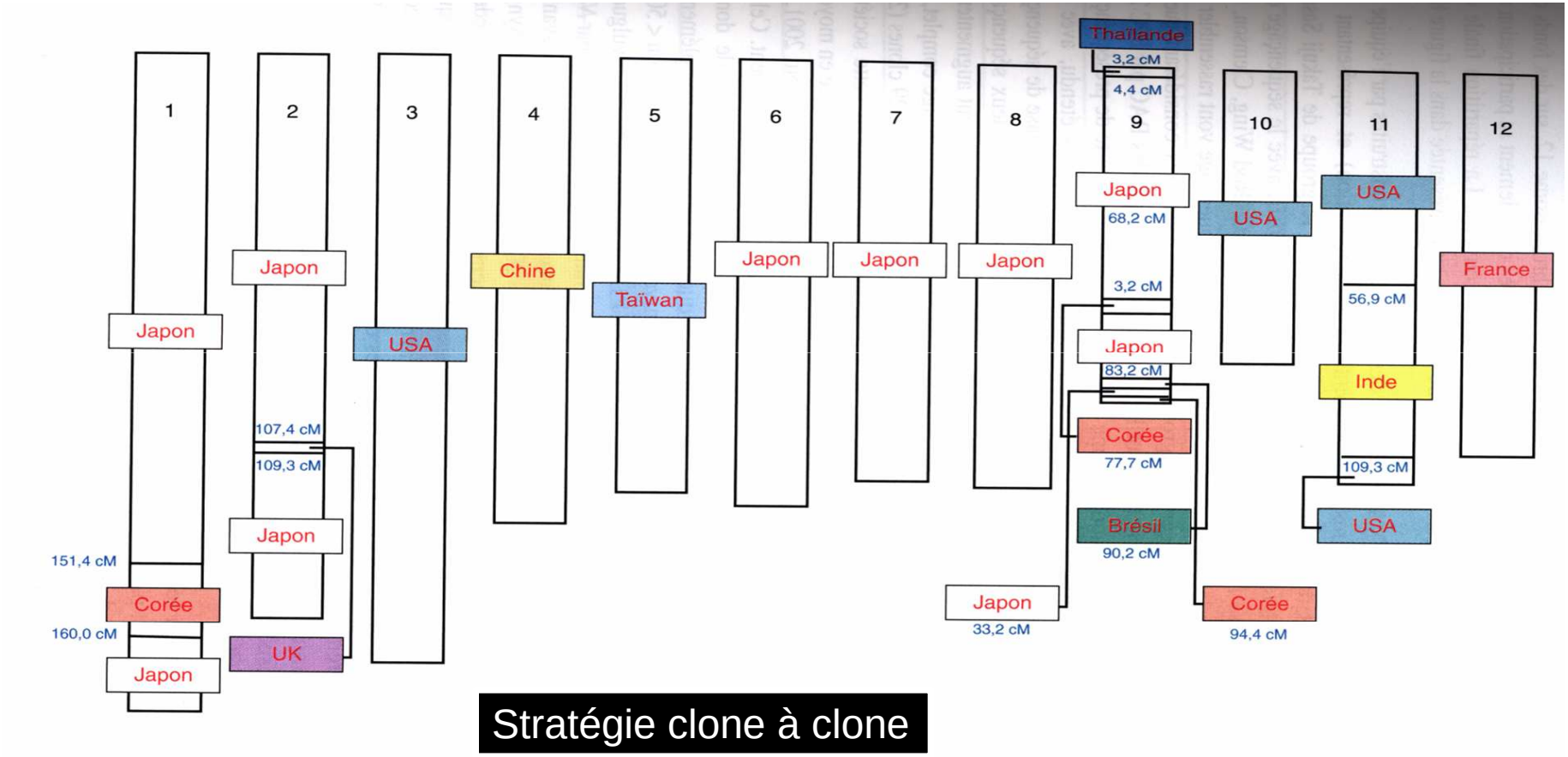


Figure 8. État en novembre 2002 de la répartition du séquençage des 12 chromosomes du riz entre les différents pays membres de l'IRGSP.

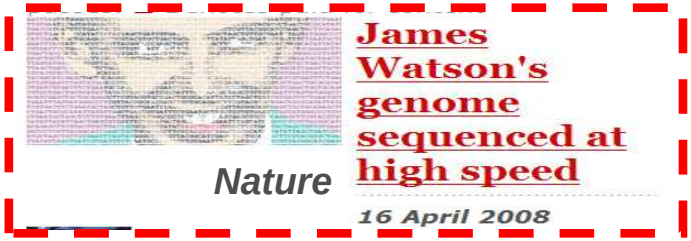
Le séquençage du Génome de Watson , n'est pas le 1er génome humain séquencé (C. Venter en 2007) mais …

nature Vol 452|17 April 2008|doi:10.1038/nature06884

LETTERS

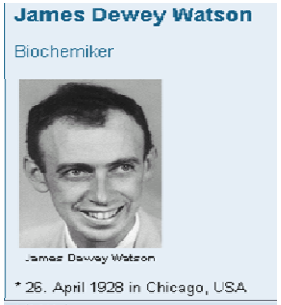
The complete genome of an individual by massively parallel DNA sequencing

David A. Wheeler[1]*, Maithreyan Srinivasan[2]*, Michael Egholm[2]*, Yufeng Shen[1]*, Lei Chen[1], Amy McGuire[3], Wen He[2], Yi-Ju Chen[2], Vinod Makhijani[2], G. Thomas Roth[2], Xavier Gomes[2], Karrie Tartaro[2]†, Faheem Niazi[2], Cynthia L. Turcotte[2], Gerard P. Irzyk[2], James R. Lupski[4,5,6], Craig Chinault[4], Xing-zhi Song[1], Yue Liu[1], Ye Yuan[1], Lynne Nazareth[1], Xiang Qin[1], Donna M. Muzny[1], Marcel Margulies[2], George M. Weinstock[1,4], Richard A. Gibbs[1,4] & Jonathan M. Rothberg[2]†



Stratégie UHTS

J. Craig Venter



QUICKER, SMALLER, CHEAPER

Genome sequenced (publication year)	HGP (2003)	Venter (2007)	Watson (2008)
Time taken (start to finish)	13 years	4 years	4.5 months
Number of scientists listed as authors	> 2,800	31	27
Cost of sequencing (start to finish)	$2.7 billion	$100 million	< $1.5 million
Coverage	8–10 ×	7.5 ×	7.4 ×
Number of institutes involved	16	5	2
Number of countries involved	6	3	1

Human Genome Project

6 Nov 2008

Current Issue
Volume 455
Number 7218

This week's news

- What Obama's win means for science
- HIV vaccine failure explained?
- Blink and you'll miss it
- More from Nature News

Latest Research

Vol 456|6 November 2008|doi:10.1038/nature07517 nature

ARTICLES

Accurate whole human genome sequencing using reversible terminator chemistry

A list of authors and their affiliations appears at the end of the paper

nature Vol 456|6 November 2008|doi:10.1038/nature07485

ARTICLES

DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome

Timothy J. Ley[1,2,3,4]*, Elaine R. Mardis[2,3]*, Li Ding[2,3], Bob Fulton[3], Michael D. McLellan[3], Ken Chen[3], David Dooling[3], Brian H. Dunford-Shore[3], Sean McGrath[3], Matthew Hickenbotham[3], Lisa Cook[3], Rachel Abbott[3], David E. Larson[3], Dan C. Koboldt[3], Craig Pohl[3], Scott Smith[3], Amy Hawkins[3], Scott Abbott[3], Devin Locke[3], LaDeana W. Hillier[3,8], Tracie Miner[3], Lucinda Fulton[3], Vincent Magrini[2,3], Todd Wylie[3], Jarret Glasscock[3], Joshua Conyers[3], Nathan Sander[3], Xiaoqi Shi[3], John R. Osborne[3], Patrick Minx[3], David Gordon[8], Asif Chinwalla[3], Yu Zhao[1], Rhonda E. Ries[1], Jacqueline E. Payton[5], Peter Westervelt[1,4], Michael H. Tomasson[1,4], Mark Watson[3,4,5], Jack Baty[6], Jennifer Ivanovich[4,7], Sharon Heath[1,4], William D. Shannon[1,4], Rakesh Nagarajan[4,5], Matthew J. Walter[1,4], Daniel C. Link[1,4], Timothy A. Graubert[1,4], John F. DiPersio[1,4] & Richard K. Wilson[2,3,4]

Acute myeloid leukaemia is a highly malignant haematopoietic tumour that affects about 13,000 adults in the United States each year. The treatment of this disease has changed little in the past two decades, because most of the genetic events that initiate the disease remain undiscovered. Whole-genome sequencing is now possible at a reasonable cost and timeframe to use this approach for the unbiased discovery of tumour-specific somatic mutations that alter the protein-coding genes. Here we present the results obtained from sequencing a typical acute myeloid leukaemia genome, and its matched normal counterpart obtained from the same patient's skin. We discovered ten genes with acquired mutations; two were previously described mutations that are thought to contribute to tumour progression, and eight were new mutations present in virtually all tumour cells at presentation and relapse, the function of which is not yet known. Our study establishes whole-genome sequencing as an unbiased method for discovering cancer-initiating mutations in previously unidentified genes that may respond to targeted therapies.

nature Vol 456|6 November 2008|doi:10.1038/nature07484

ARTICLES

The diploid genome sequence of an Asian individual

Jun Wang[1,2,3,4]*, Wei Wang[1,3]*, Ruiqiang Li[1,3,4]*, Yingrui Li[1,5,6]*, Geng Tian[1,7], Laurie Goodman[1], Wei Fan[1], Junqing Zhang[1], Jun Li[1], Juanbin Zhang[1], Yiran Guo[1,7], Binxiao Feng[1], Heng Li[1,8], Yao Lu[1], Xiaodong Fang[1], Huiqing Liang[1], Zhenglin Du[1], Dong Li[1], Yiqing Zhao[1,7], Yujie Hu[1,7], Zhenzhen Yang[1], Hancheng Zheng[1], Ines Hellmann[9], Michael Inouye[8], John Pool[9], Xin Yi[1,7], Jing Zhao[1], Jinjie Duan[1], Yan Zhou[1], Junjie Qin[1,7], Lijia Ma[1,7], Guoqing Li[1], Zhentao Yang[1], Guojie Zhang[1,7], Bin Yang[1], Chang Yu[1], Fang Liang[1,7], Wenjie Li[1], Shaochuan Li[1], Dawei Li[1], Peixiang Ni[1], Jue Ruan[1,7], Qibin Li[1,7], Hongmei Zhu[1], Dongyuan Liu[1], Zhike Lu[1], Ning Li[1,7], Guangwu Gu[1], Jianguo Zhang[1], Jia Ye[1], Lin Fang[1], Qin Hao[1,7], Quan Chen[1,5], Yu Liang[1,7], Yeyang Su[1,7], A. san[1,7], Cuo Ping[1,7], Shuang Yang[1], Fang Chen[1,7], Li Li[1], Ke Zhou[1], Hongkun Zheng[1,4], Yuanyuan Ren[1], Ling Yang[1], Yang Gao[1,6], Guohua Yang[1,2], Zhuo Li[1], Xiaoli Feng[1], Karsten Kristiansen[4], Gane Ka-Shu Wong[1,10], Rasmus Nielsen[9], Richard Durbin[8], Lars Bolund[1,11], Xiuqing Zhang[1,6], Songgang Li[1,2,5], Huanming Yang[1,2,3] & Jian Wang[1,2,3]

Here we present the first diploid genome sequence of an Asian individual. The genome was sequenced to 36-fold ave coverage using massively parallel sequencing technology. We aligned the short reads onto the NCBI human referenc genome to 99.97% coverage, and guided by the reference genome, we used uniquely mapped reads to assemble a high-quality consensus sequence for 92% of the Asian individual's genome. We identified approximately 3 million single-nucleotide polymorphisms (SNPs) inside this region, of which 13.6% were not in the dbSNP database. Genoty analysis showed that SNP identification had high accuracy and consistency, indicating the high sequence quality of t assembly. We also carried out heterozygote phasing and haplotype prediction against HapMap CHB and JPT haploty (Chinese and Japanese, respectively), sequence comparison with the two available individual genomes (J. D. Watson a C. Venter), and structural variation identification. These variations were considered for their potential biological impact sequence data and analyses demonstrate the potential usefulness of next-generation sequencing technologies for per genomics.

NATURE|Vol 456|20 November 2008

DNA SEQUENCING

Mammoth genomics

Michael Hofreiter

Reconstruction of most of the genome sequence of the woolly mammoth illustrates how such investigations will pave the way for a deeper understanding of the biology and evolution of extinct species.

En décembre 2011 :

Génomes séquencés et en cours de séquençage :

8847 procaryotes	2987 virus	1518 Eucaryotes
29 groupes	15 groupes	561 animaux 449 champignons 235 protistes 259 plantes

Le projet HapMap :

- comparer les séquences génétiques de différents individus afin de relever les régions chromosomiques où des variations génétiques sont partagées

- découvrir les gènes qui jouent un rôle dans la santé, la maladie et la réponse des individus aux médicaments et aux facteurs environnementaux.

- information sera dans le domaine public,

collaboration entre
Japon, du Royaume-Uni, du Canada, de la Chine, du Nigeria et des États-Unis.

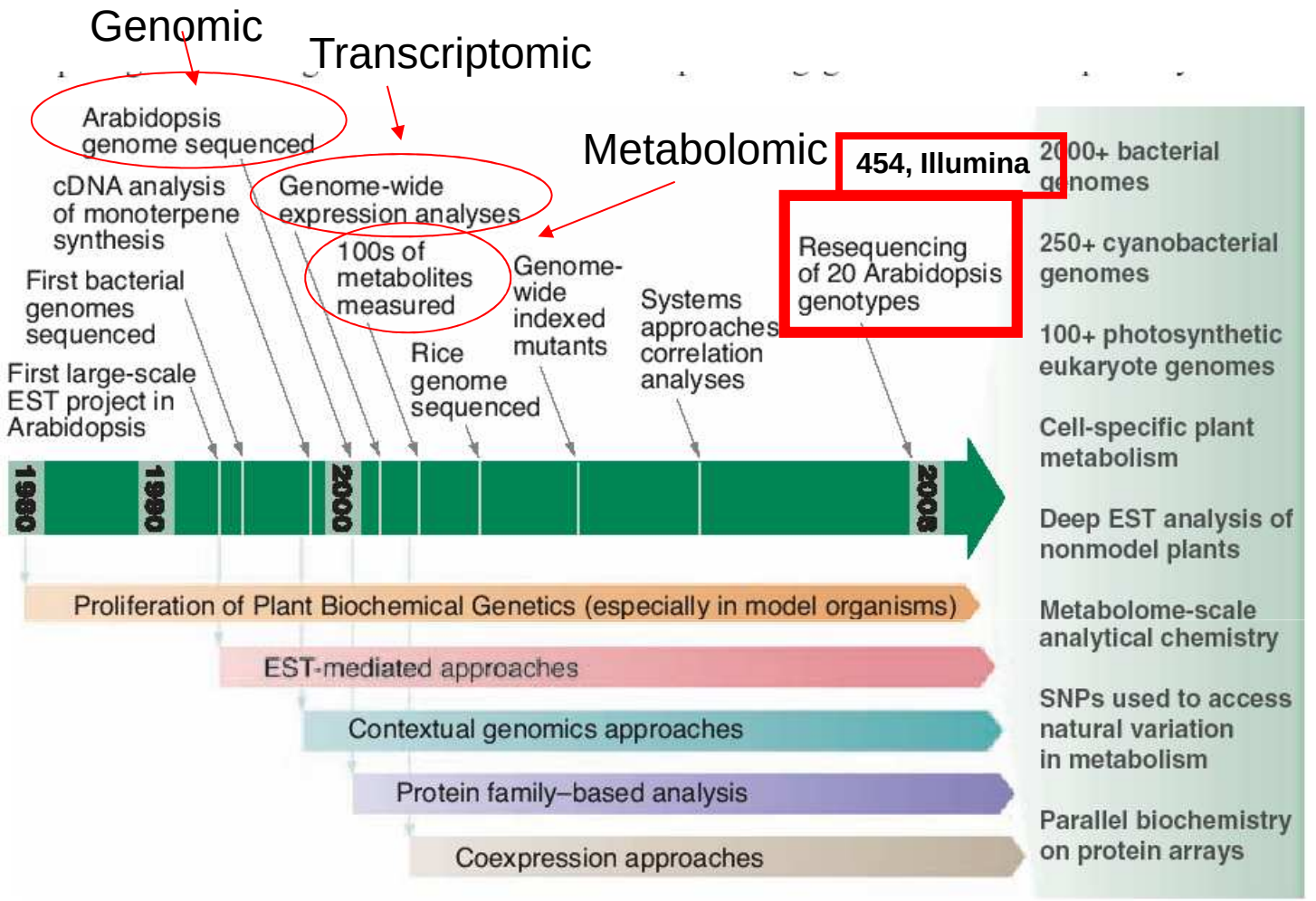
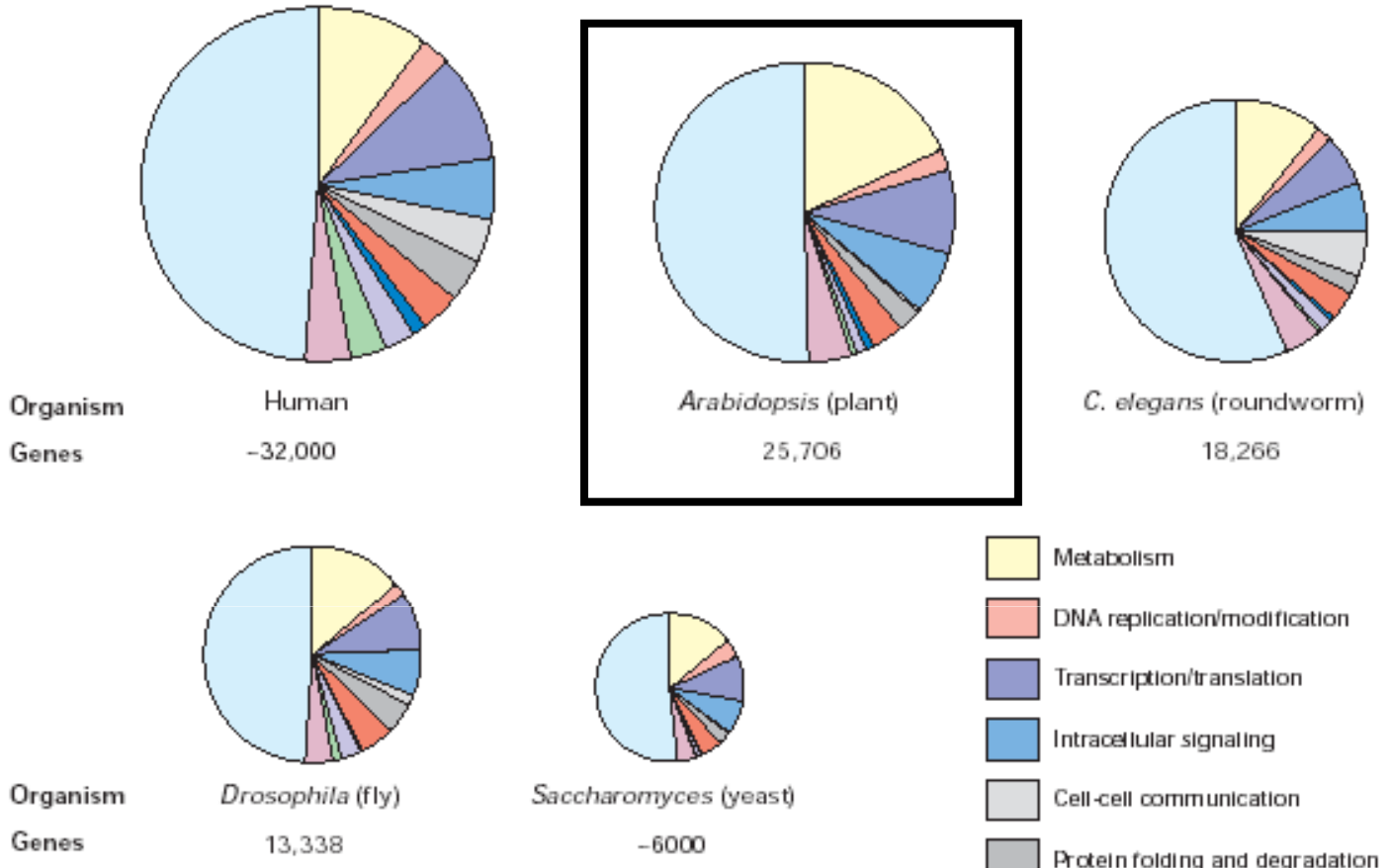


Fig. 1. Time line of genomics-enabled plant biochemistry. Selected major advances to date are indicated above and below the time line. Some approaches and tools likely to further understanding of plant biochemistry during the coming decade are indicated to the right of the time line.

Rappel UHTS

Comparaison des génomes



▲ **FIGURE 9-34 Comparison of the number and types of proteins encoded in the genomes of different eukaryotes.** For each organism, the area of the entire pie chart represents the total number of protein-coding genes, all shown at roughly the same scale. In most cases, the functions of the proteins encoded by about half the genes are still unknown (light blue). The functions of the remainder are known or have been predicted by sequence similarity to genes of known function. [Adapted from International Human Genome Sequencing Consortium, 2001, *Nature* **409**:860.]

De la connaissance du génome…

- **A son annotation …**
- A son étude de l’expression …
- Vers l’analyse fonctionnelle…

Yamada et al. *Science*, 302, 842-846 , 2003. Empirical Analysis of Transcriptional Activity in the *Arabidopsis* Genome

Arabidospsis : 26828 gènes prédits dont 25540 prédits comme gènes codants

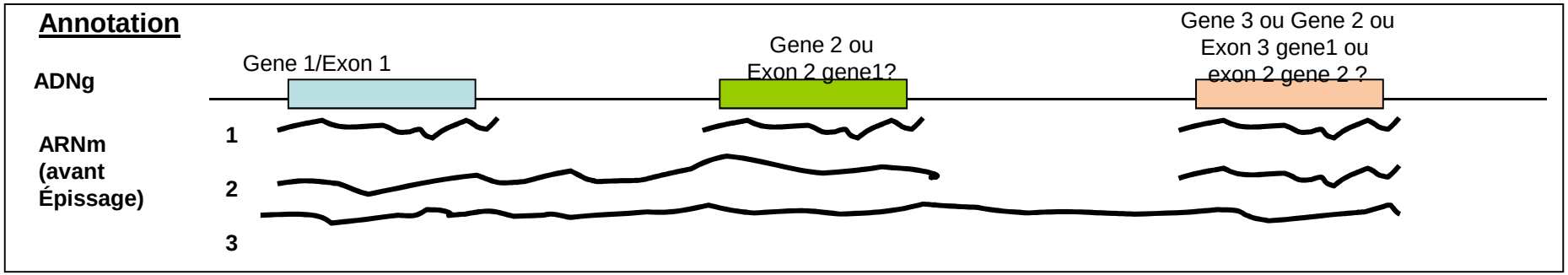
3 classes

AE
Annotated
expressed
70 % (ESTs)

ANE
Annotated
non
Expressed 30%

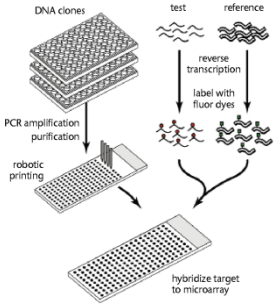
NAE
Non
Annotated
expressed ?

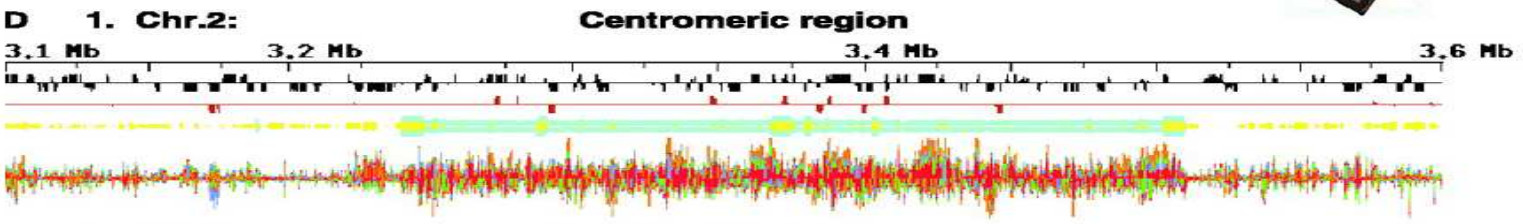
→intérêt des collections d'EST et de cDNA pleine taille (full lenght)



Solution ? **High density oligo arrays qui couvre 94 % génome Arabidopsis**

Soit 12 arrays avec 834 000 oligos de 25 mers / arrays
Hybridation avec 4 populations d'ARNm





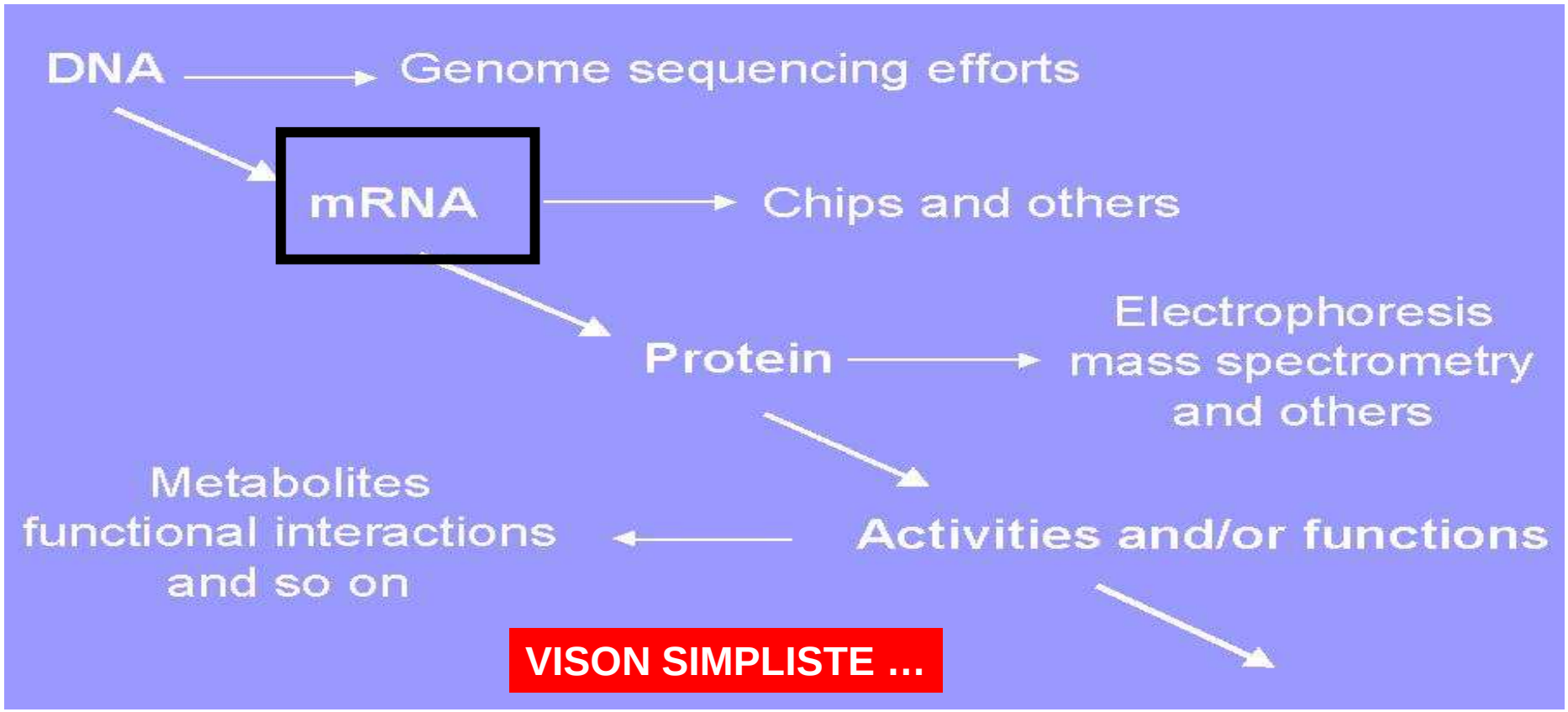
Detection d'activité transcriptionnelle dans région centromérique (40 gènes) et dans régions intergéniques (2000)

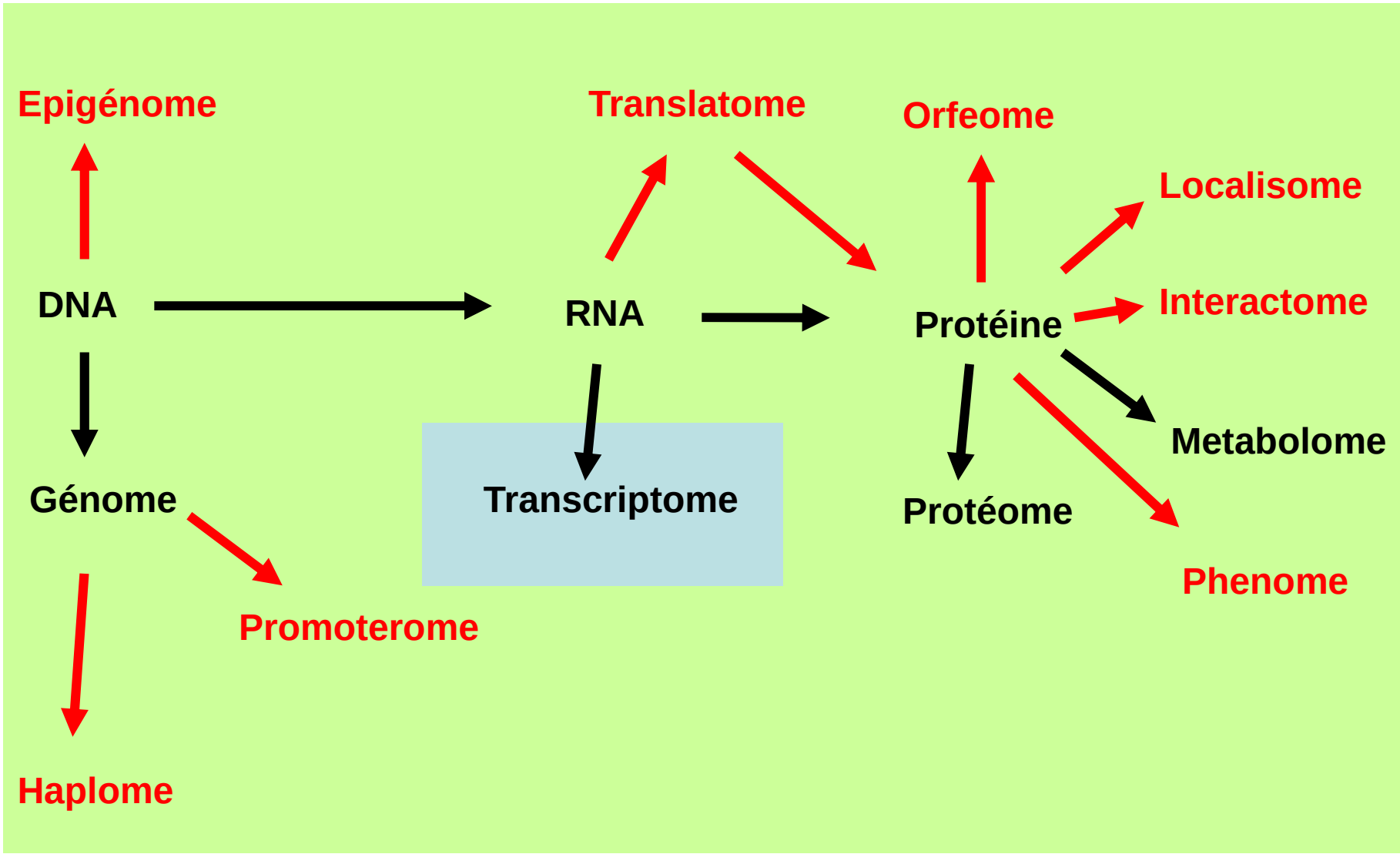
Bilan : 5817 nouvelles unités de transcription soit 30 % de plus que prédit par annotation

De la génomique à la protéomique …

De la relation : 1 gène } 1 protéine...

De la relation : 1 protéine } 1 fonction ...

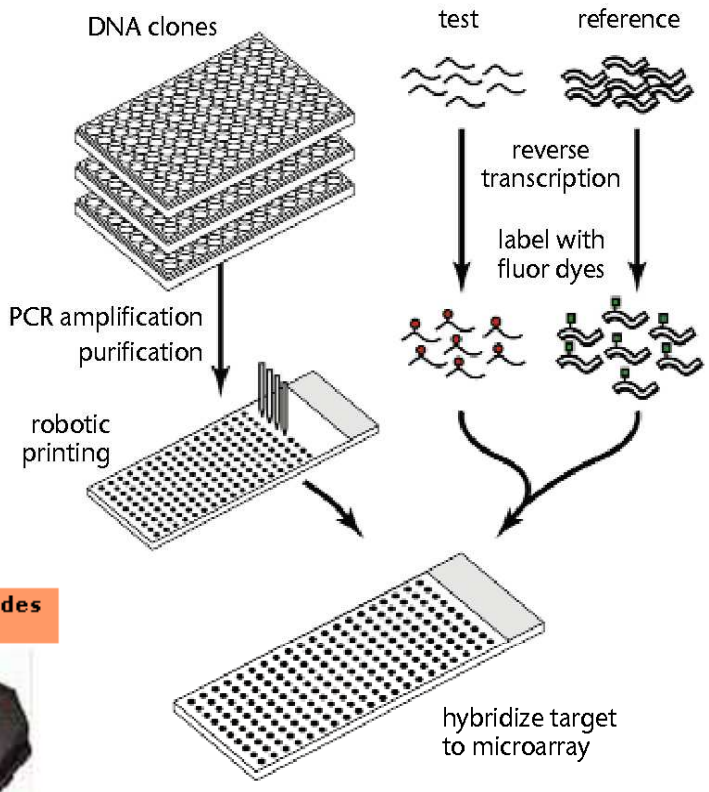


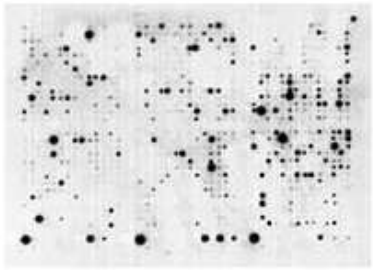
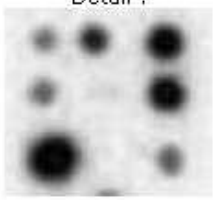
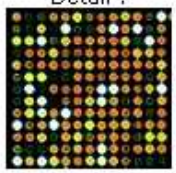

Epigénome
Translatome
Orfeome
Localisome
DNA
RNA
Protéine
Interactome
Transcriptome
Metabolome
Génome
Protéome
Promoterome
Phenome
Haplome

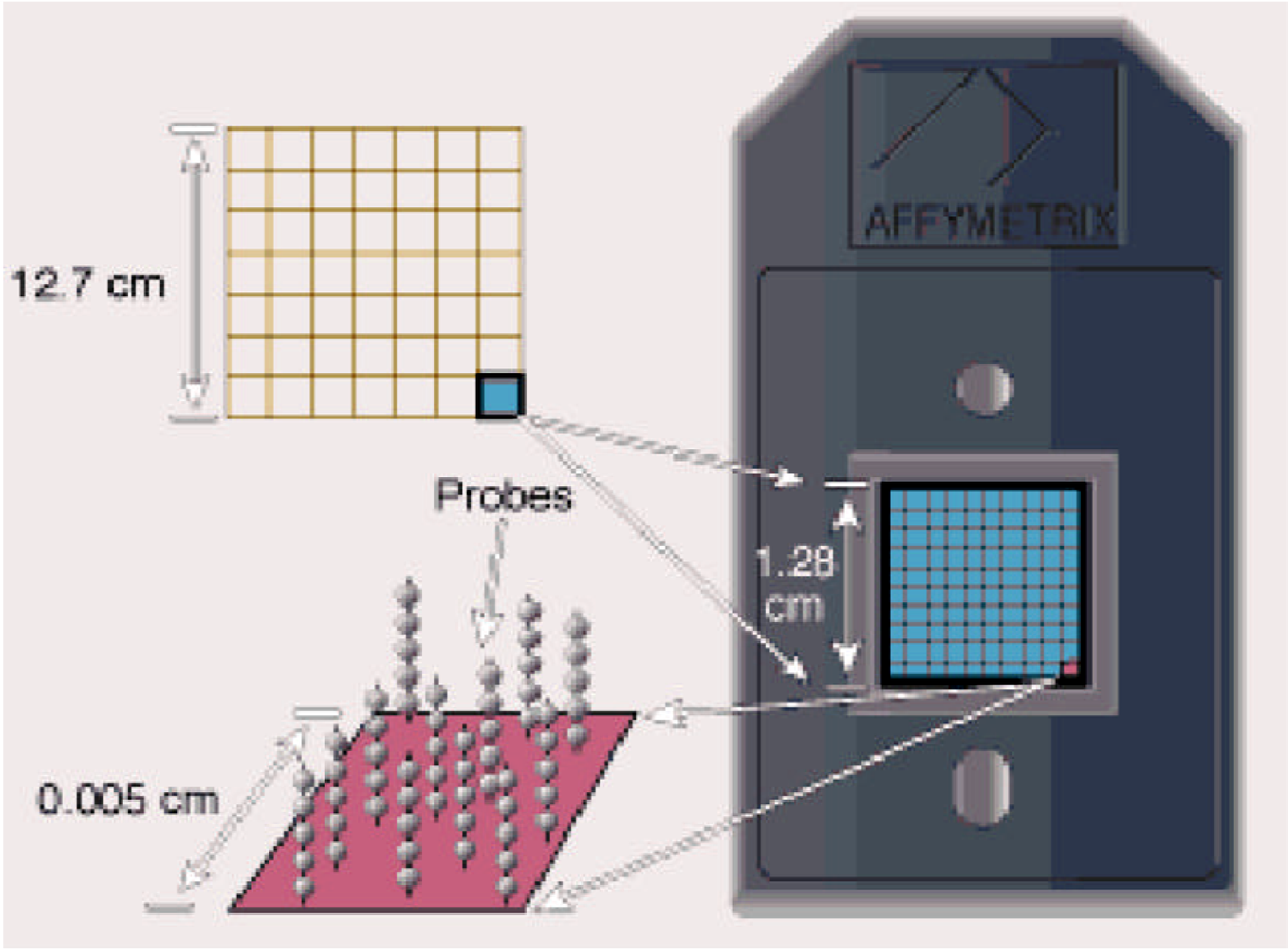
Puces à ADN :

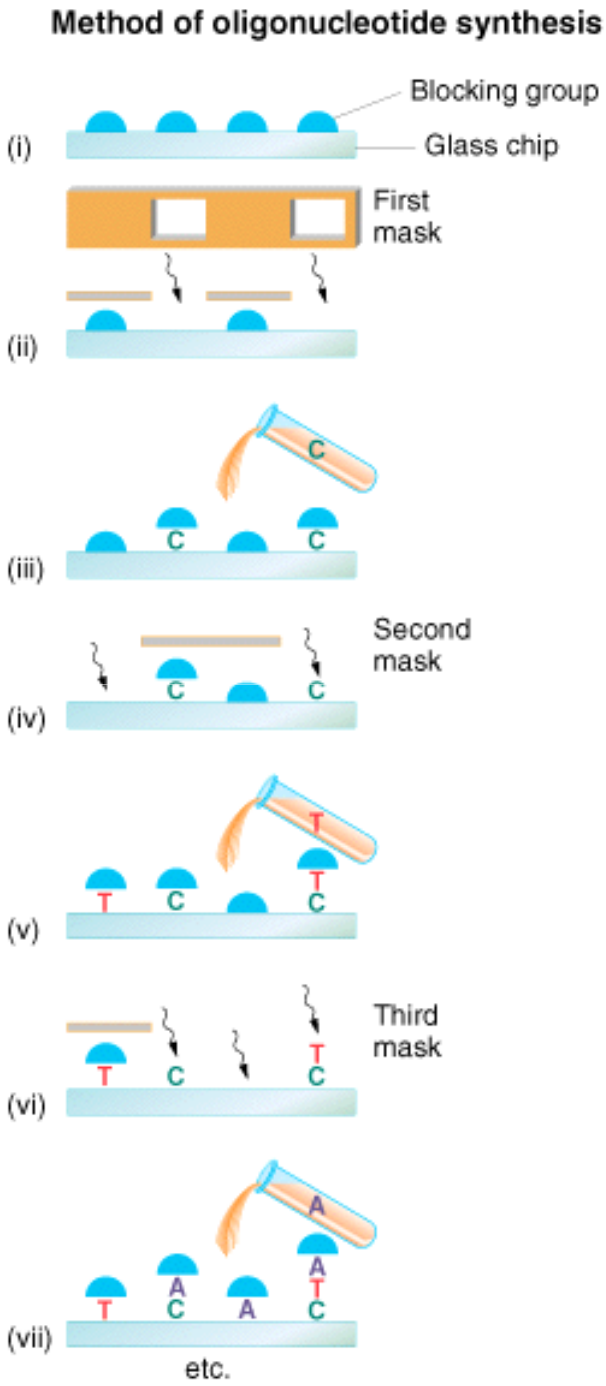
- des cDNAs à la puce ?

- Quels supports ?



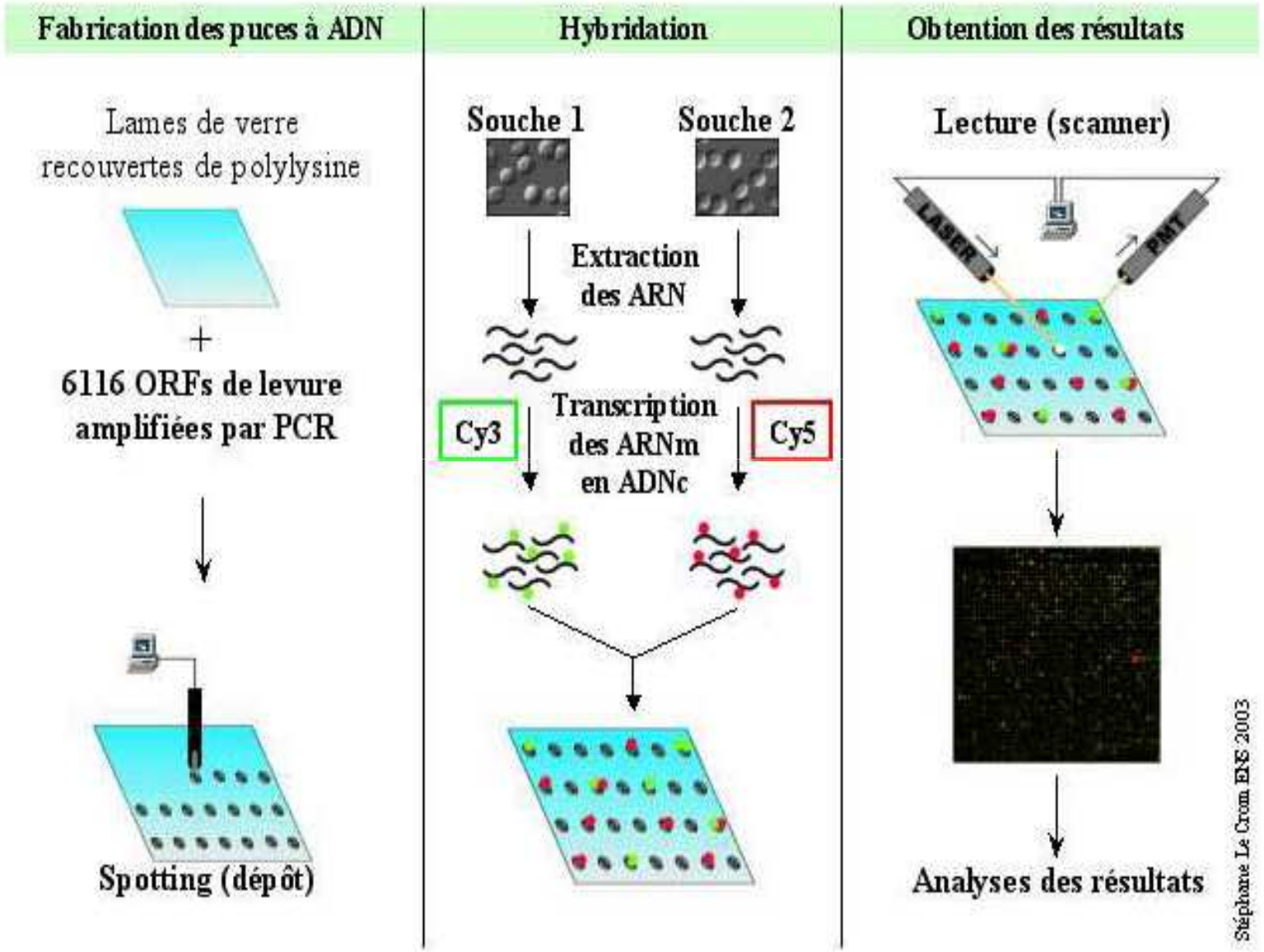
Filtres haute densité (macroarrays)	Lames de verre (microarrays)	Puces à oligonucléotides
		
Détail : 	Détail : 	Détail : 
Taille : 12cm x 8cm	Taille : 5,4cm x 0,9cm	Taille : 1,28cm x 1,28cm
◆ 2400 clones par membrane ◆ marquage radioactif ◆ 1 condition expérimentale par membrane	◆ 10000 clones par lame ◆ marquage fluorescent ◆ 2 conditions expérimentales par lame	◆ 300000 oligonucléotides par lame ◆ marquage fluorescent ◆ 1 condition expérimentale par lame


12.7 cm
Probes
1.28 cm
0.005 cm
AFFYMETRIX



Synthesis of an oligonucleotide gene chips - part I

Puces à ADN : Principe



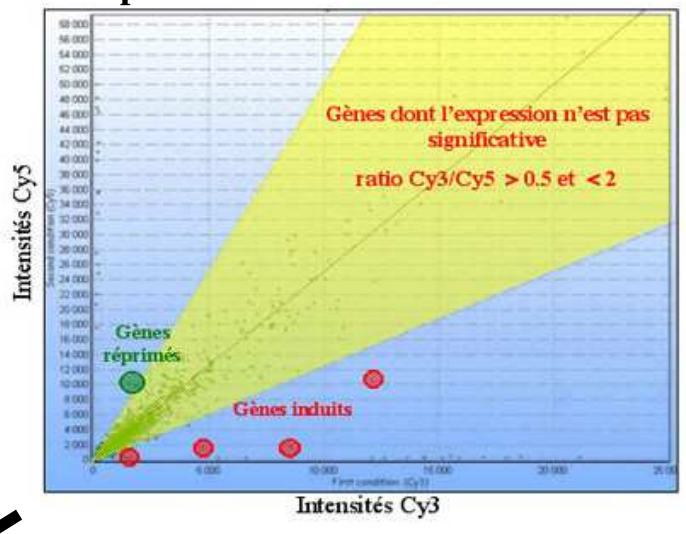
Puces à ADN: du résultat à la signification biologique ?

Rouge : Spé A
Vert : Spé B
Jaune : Commun

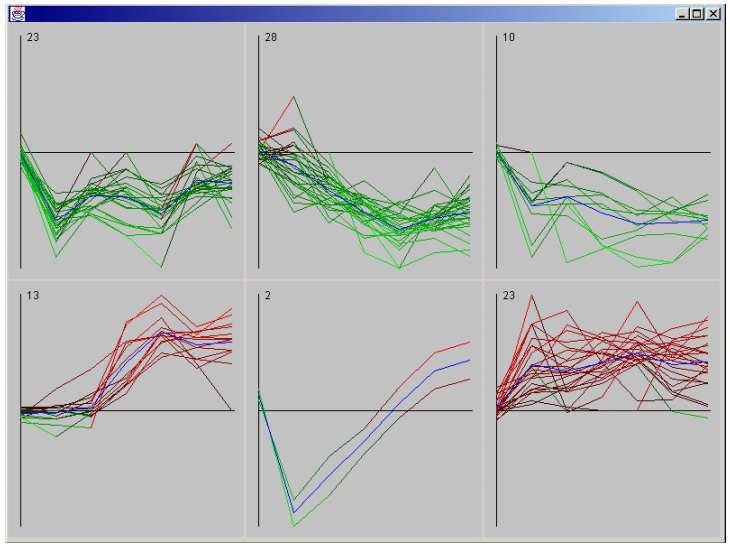
Sonde A

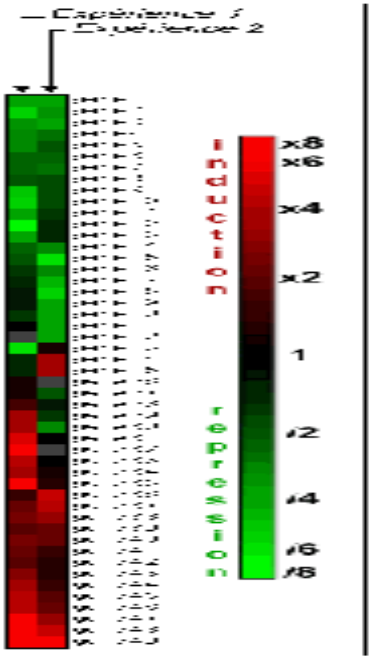
Sonde B

Quels sont les gènes dont l'expression est réprimée ? Induite ? Invariable ?



Quels sont les gènes dont le profil d'expression est identique ?





Le clustering d'expression

MPSS : ?
Massively Parallel Signature Sequencing

UHTS method

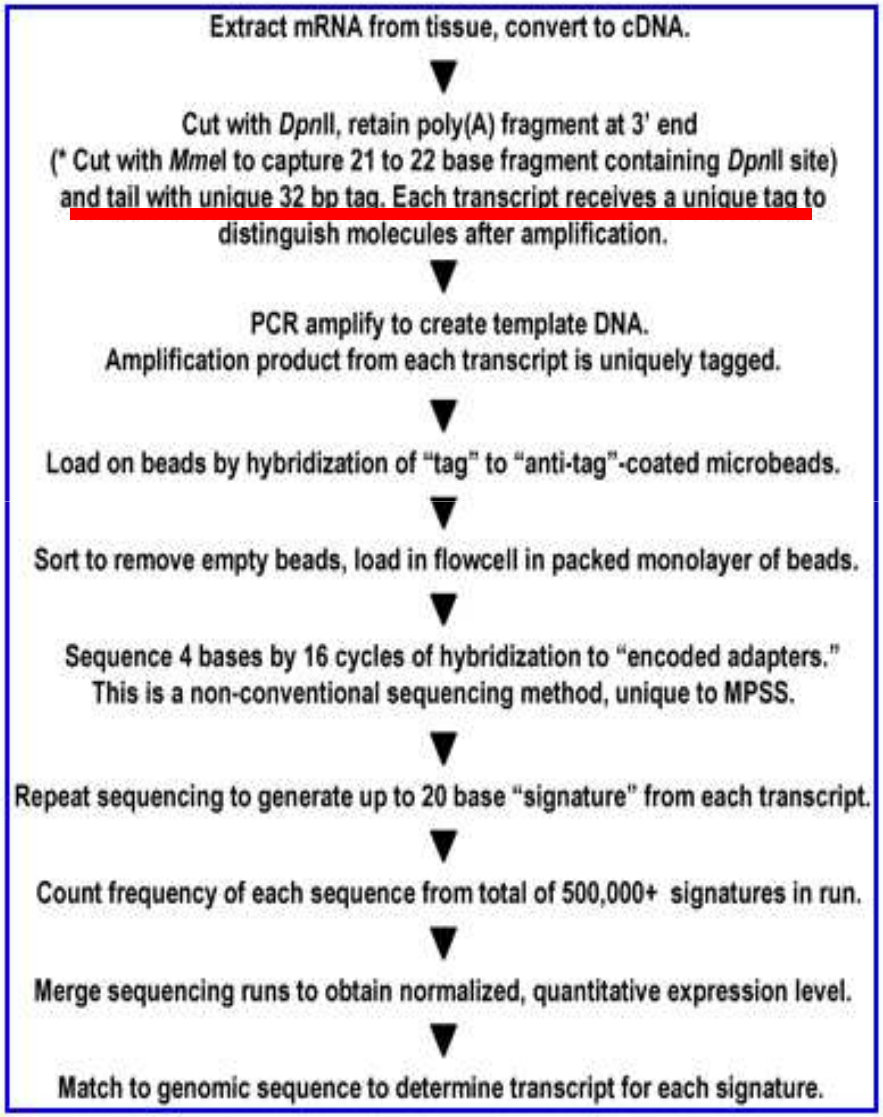
Genome Research 14:1641-1653, 2004
ISSN 1088-9051/ $5.00

Methods

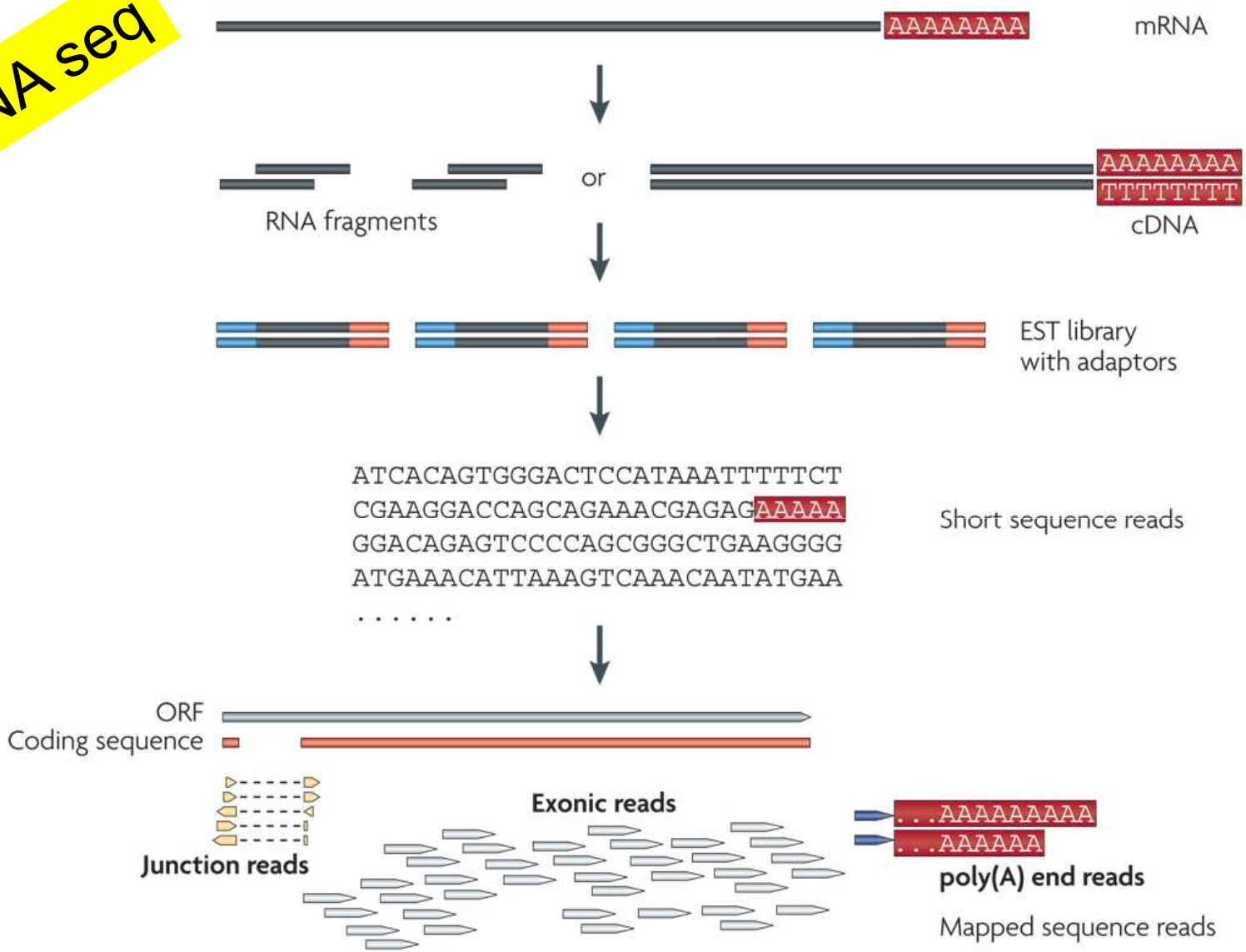
The Use of MPSS for Whole-Genome Transcriptional Analysis in *Arabidopsis*

Blake C. Meyers[1,4], Shivakundan Singh Tej[1], Tam H. Vu[1], Christian D. Haudenschi Vikas Agrawal[1], Steve B. Edberg[2], Hassan Ghazal[1] and Shannon Decola[3]

[1] *Department of Plant and Soil Sciences, and Delaware Biotechnology Institute, Univer Delaware, Newark, Delaware 19714, USA;* [2] *Department of Vegetable Crops, Universi California, Davis, California 95616, USA;* [3] *Lynx Therapeutics, Inc., Hayward, Califor 94545, USA*



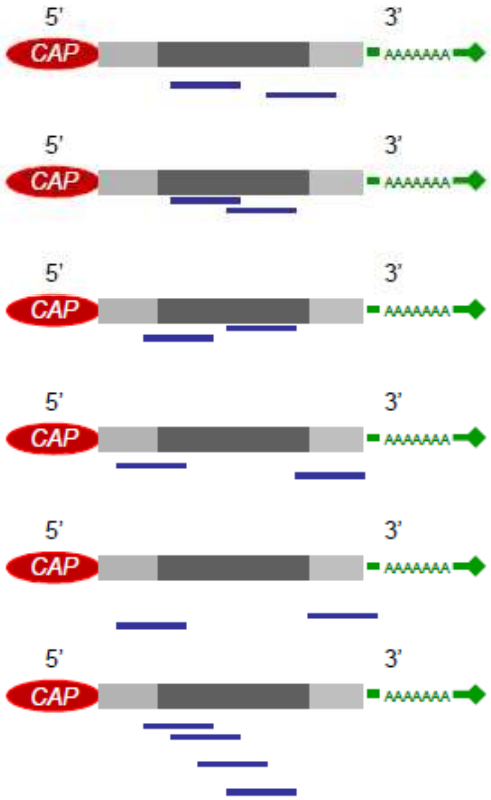
RNA seq


AAAAAAAA
mRNA
or
RNA fragments
AAAAAAAA
TTTTTTTT
cDNA
EST library
with adaptors
ATCACAGTGGGACTCCATAAATTTTTCT
CGAAGGACCAGCAGAAACGAGAGAAAAA
GGACAGAGTCCCCAGCGGGCTGAAGGGG
ATGAAACATTAAAGTCAAACAATATGAA
. . . . . .
Short sequence reads
ORF
Coding sequence
Exonic reads
Junction reads
...AAAAAAAAA
...AAAAAA
poly(A) end reads
Mapped sequence reads

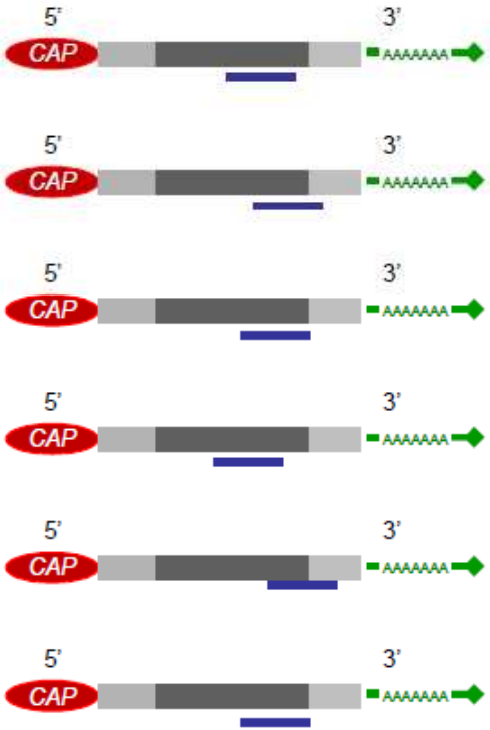
RNA seq

Resequencing Strategy Summary

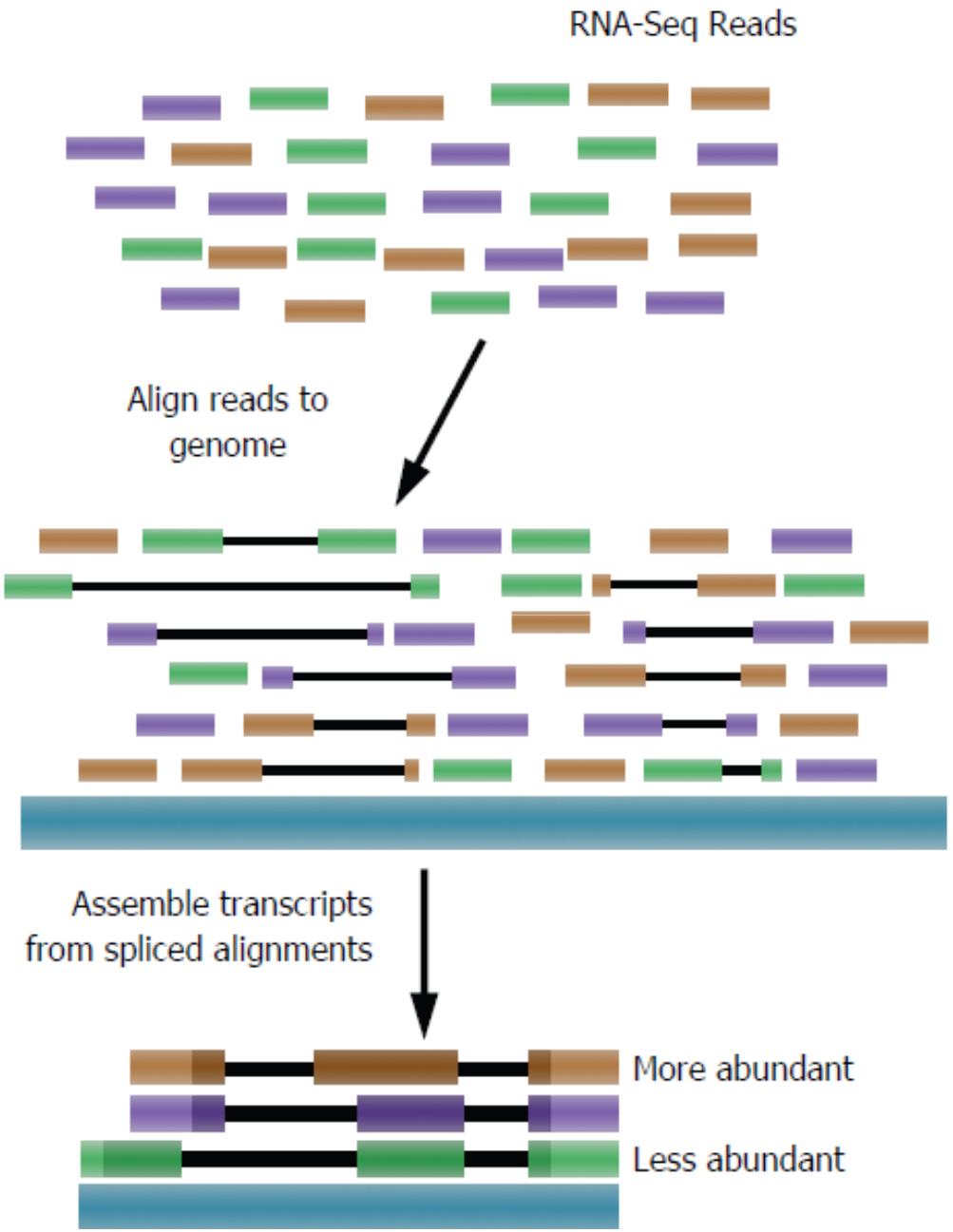
RNA seq



DGE



RNA seq


RNA-Seq Reads
Align reads to genome
Assemble transcripts from spliced alignments
More abundant
Less abundant

ChIP on chip

Chromatin ImmunoPrecipitation on Chip

ChIP and sequencing high throughput

utilisées pour repérer des sites de fixation de facteurs de transcription

localiser ces sites et d'étudier les séquences d'ADN correspondantes

1.Liaison covalente in vivo des protéines à l'ADN
2.Extraction de l'ADN
3.Découpage de l'ADN par sonication
4.Sélection des fragments grâce à un anticorps
5.Précipitation des complexes ADN-protéine-anticorps,
6. Séparation du complexe ADN-protéine pour garder ADN (proteinase K)
7. On obtient une collection de fragments d'ADN qui interagissent avec une protéine d'intérêt .

combinaison de la technique de Chromatin Immunoprécipitation avec la méthode des puces à ADN.

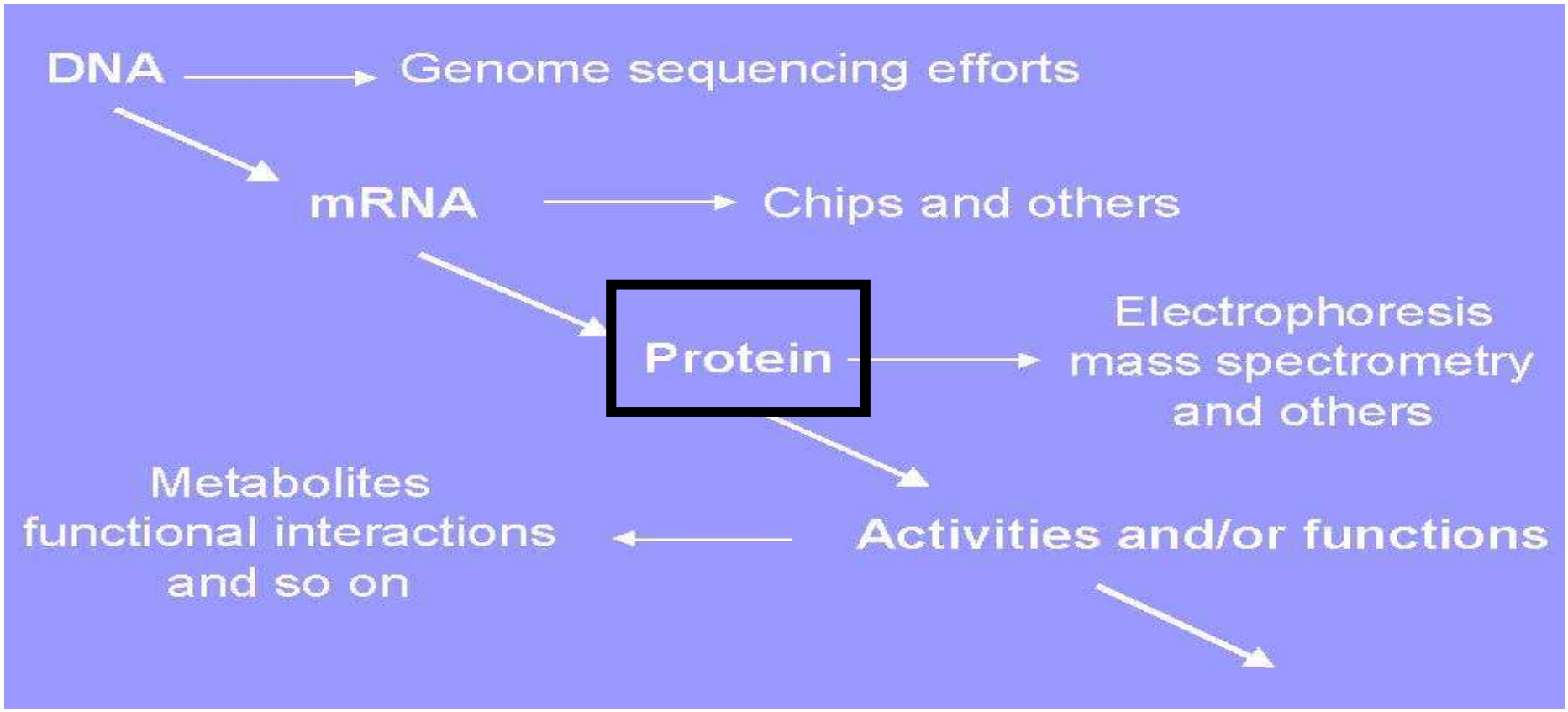
8.On utilise une puce à ADN pour identifier les fragments

8.On séquence (illumina)… pour identifier les fragments

De la *génomique* à la protéomique …

De la relation : 1 *gène* } 1 protéine...

De la relation : 1 protéine } 1 fonction ...

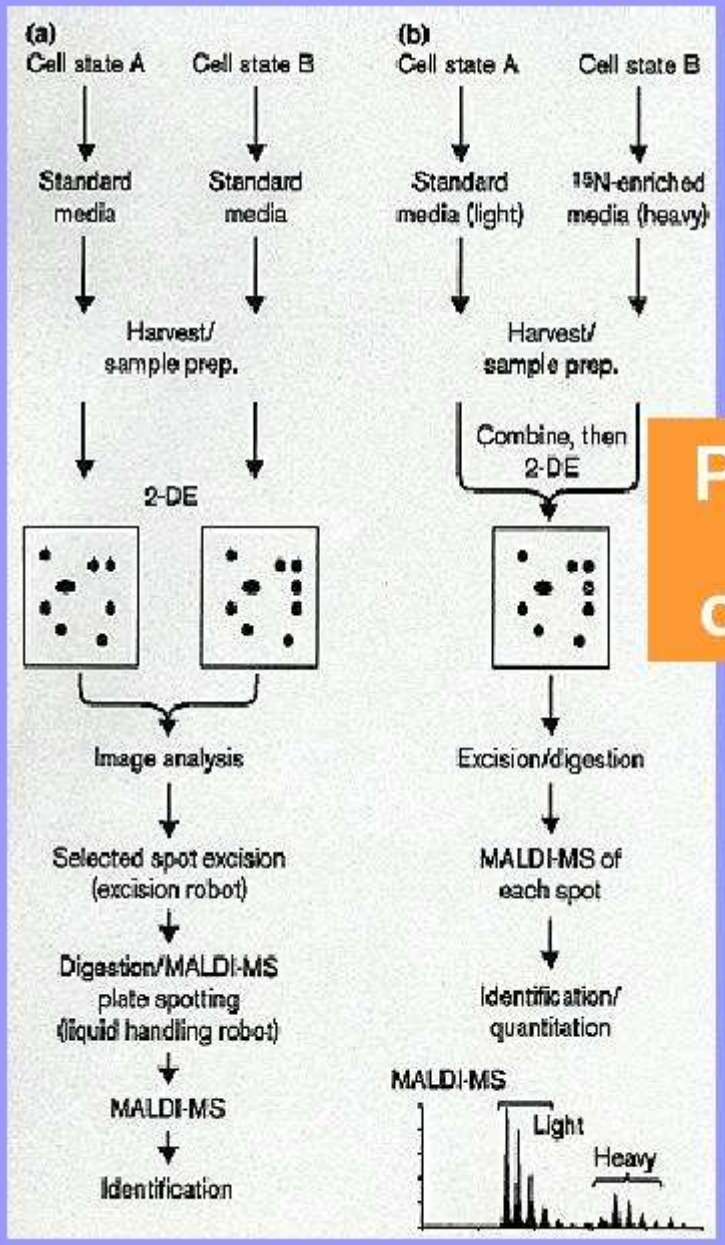


Defining Proteomics

- Proteomics - in the post-genomic era
  - Differential display of proteins on 2-dimensional gels
    - Application to mutants, knockouts, different stimuli, etc.
  - Protein identification
    - Affinity purification (analysis on 1-D or 2-D gels)
    - Proteins chips (coated with proteins or antibodies)
  - Protein modification
    - Phosphorylation, glycosylation, ….
  - Protein-protein interaction
    - Yeast two-hybrid, phage display
    - Affinity purification of protein complexes
  - Protein identification by mass spectrometry

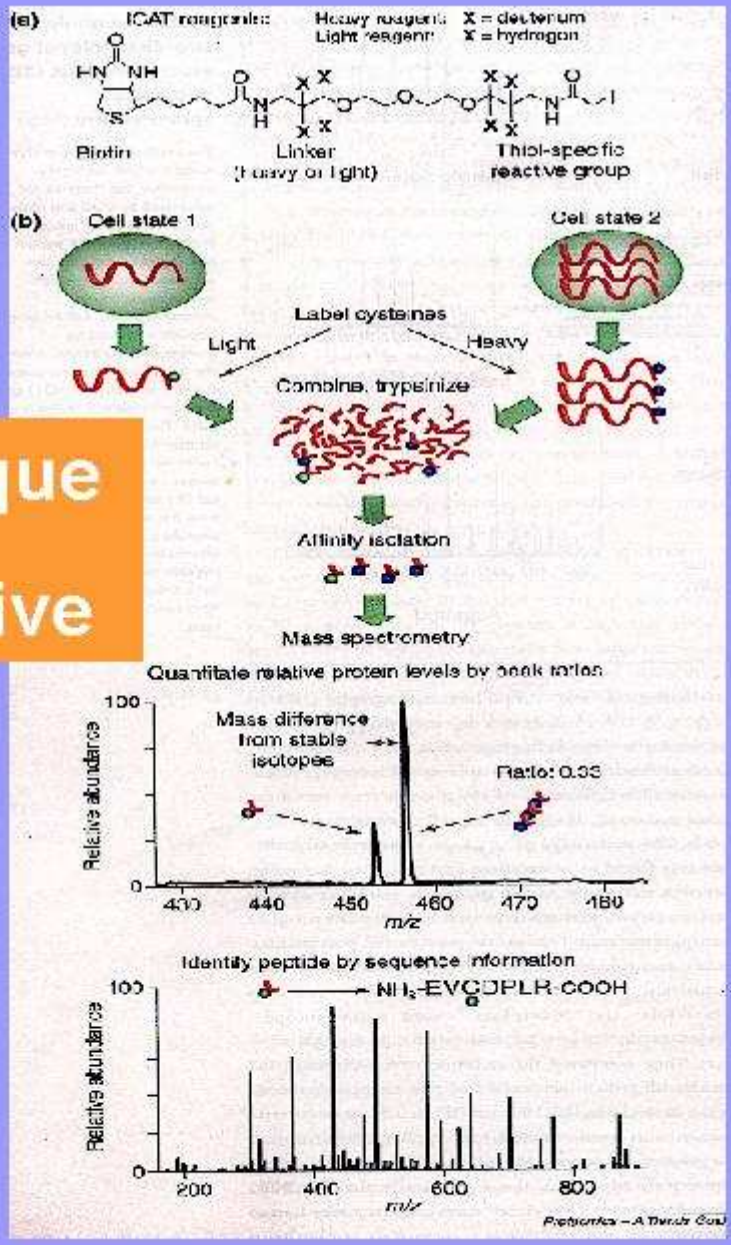
Méthode "standard"

Marquage différentiel



Protéomique comparative

Isotope-Coded Affinity Tag (ICAT)



Protéomique systématique

NATURE|Vol 452|24 April 2008

Biologists initiate plan to map human proteome

Ambitious plans to catalogue and characterize all proteins in the human body — a Human Proteome Project — are being drawn up by a small group of researchers. But with a price tag of around US$1 billion, some question whether the organizers can raise enough money or momentum for such an undertaking.

Researchers looked into the idea in the mid-1990s as the Human Genome Project was taking shape — the human proteome seemed a natural successor. However, a coordinated effort to index human proteins never emerged. One reason is that the scale and complexity of the problem proved daunting and nebulous. Protein-coding genes in the body can make tens of different versions of a protein, and each of these can be modified by the addition of chemical groups in countless different ways. All these proteins are being manufactured at differing levels, and at different moments in time, in the 200 or so types of human cell. "It was thought to be beyond comprehension," says John Bergeron of McGill University in Montreal, Canada, former president of the Human Proteome Organisation (HUPO).

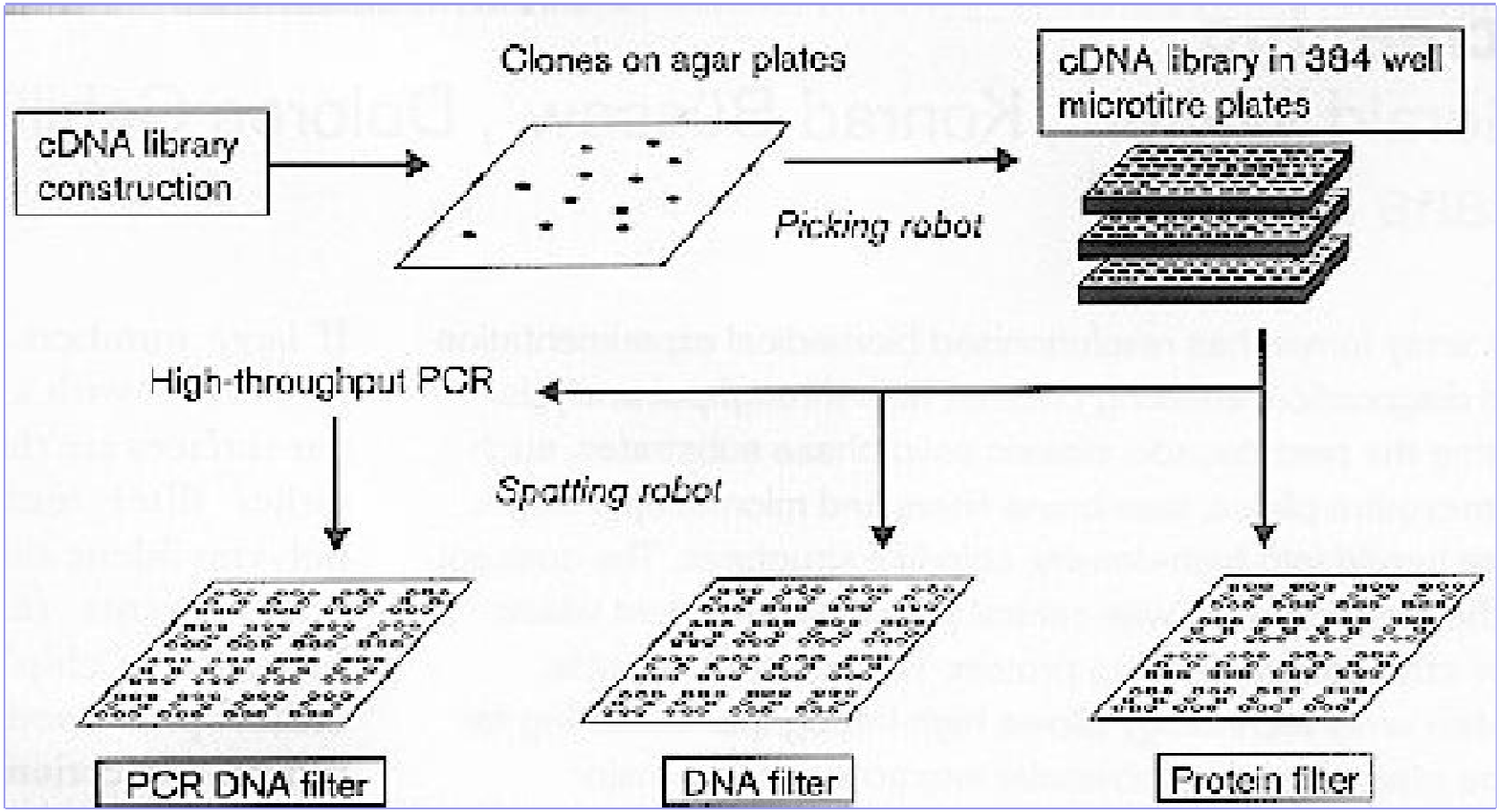
Now Bergeron and a group of leading proteomics researchers are putting together a proposal because estimates of the number of protein-coding genes have shrunk. It was once thought that there might be around 50,000 or 100,000, but now, just 21,000 or so are thought to exist, making the scale of human proteomics more manageable. And the group plans to focus on only a single protein produced from each gene, rather than its many forms. "We got rid of all this complexity," Bergeron says. "We tried to craft a project that would be doable with easy-to-track milestones."

The plan is to tackle this with three different experimental approaches. One would use mass spectrometry to identify proteins and their quantities in tissue samples; another would generate antibodies to each protein and use these to show its location in tissues and cells; and the third would systematically identify, for each protein, which others it interacts with in protein complexes. The project would also involve a massive bioinformatics effort to ensure that the data could be pooled and accessed, and the production of shared reagents.

Bergeron envisages the work being divvied up between labs around the world. He says that the first stage of the project — which involves amassing existing mass spectrometry

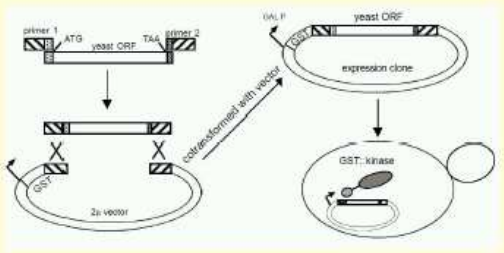
Des approches à haut débit pour la recherche de partenaires protéiques ?

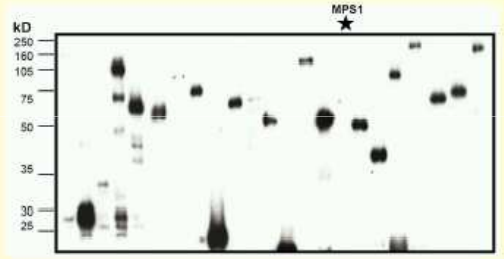
Protein arrays

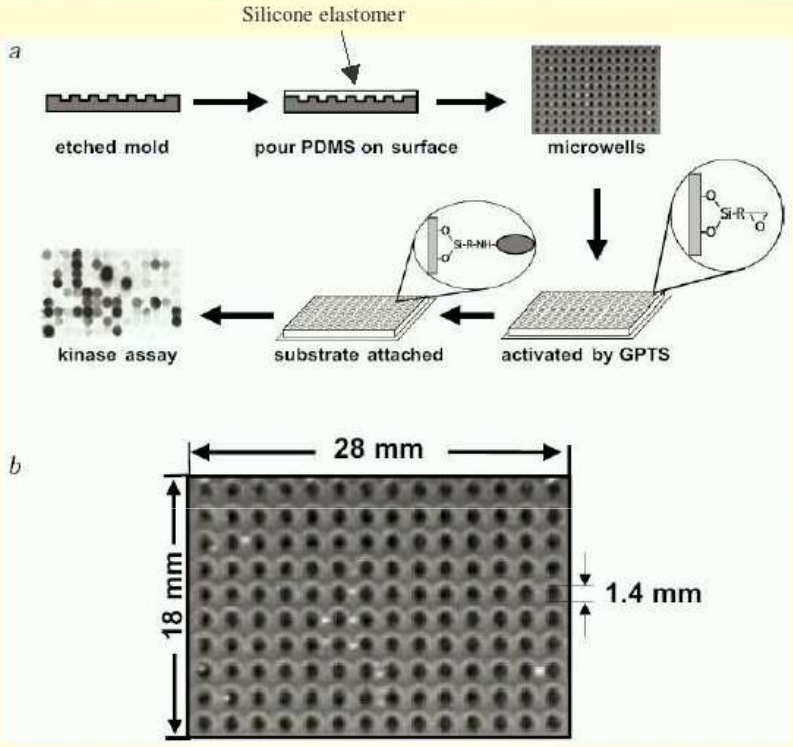


Ex: utilisation de protein chips pour la recherche de substrat d'une kinase

Analysis using protein chips







- Very good signal to noise ratio (10x better than microtiter plate)
- small amount of material needed (1/20th of what is needed for a 384-well plate
- assays are extremely sensitive (even proteins not detectable on immunblot can be assayed)
- inexpensive
- widely applicable

Nat Genet. 2000 Nov;26(3):283-9.
Analysis of yeast protein kinases using protein chips.
Zhu H, Klemic JF, Chang S, Bertone P, Casamayor A, Klemic KG, Smith D, Gerstein M, Reed MA, Snyder M.

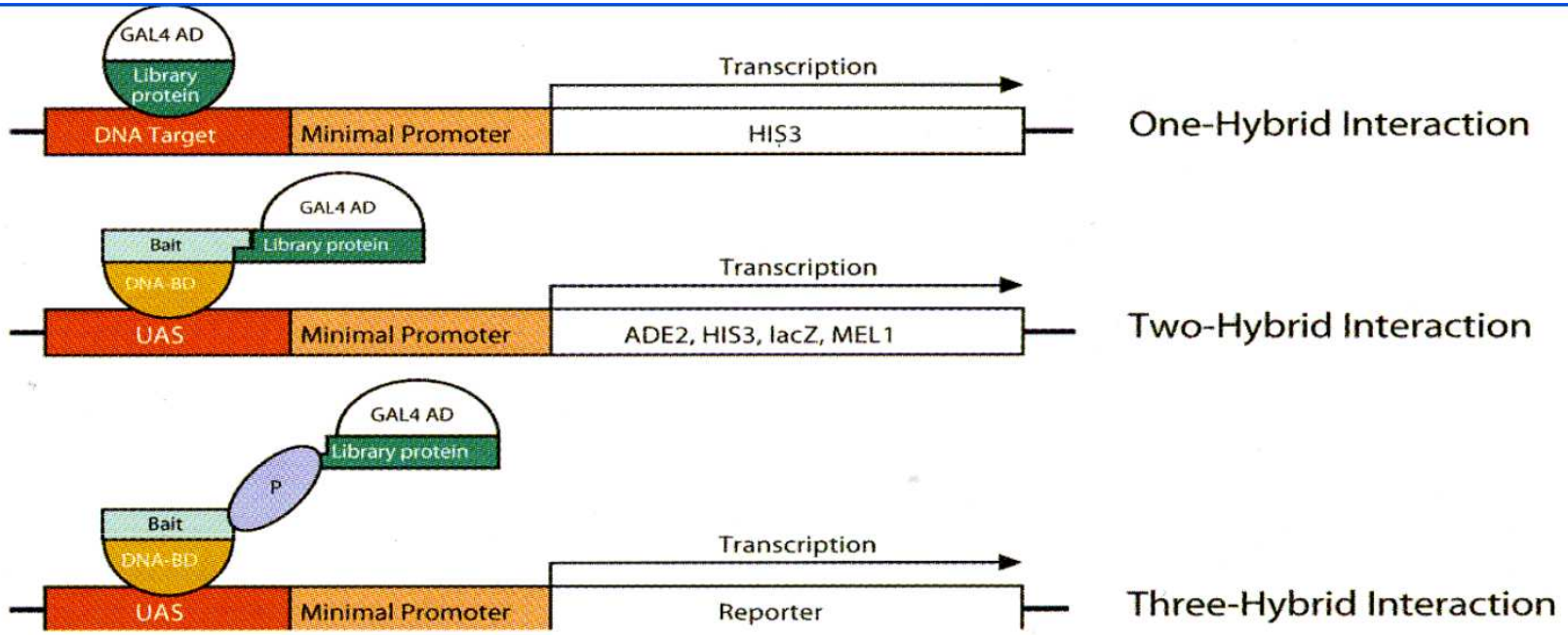
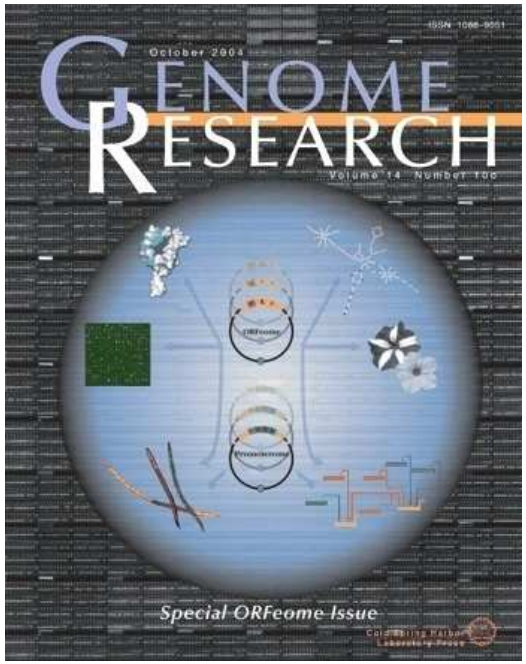


Figure 1. Yeast hybrid technology. Matchmaker systems use a sensitive transcriptional assay to detect one-hybrid, two-hybrid, and three-hybrid interactions. A yeast reporter strain transformed with both a bait and library plasmid will express the plasmids' gene inserts as fusions to either the GAL4 DNA-BD (DNA-binding domain) or AD (transcription-activating domain), depending on the plasmid (see below). If a library protein interacts with a bait protein (two-hybrid) or DNA target (one-hybrid) the host strain actively expresses the reporter gene located downstream of the promoter. Three-hybrid interactions, or protein interactions that occur via a third protein (P), can be detected using our pBridge Vector (Cat. No. 630404). UAS = GAL4-responsive upstream activating sequence.



Methods

High-Throughput Expression of *C. elegans* Proteins

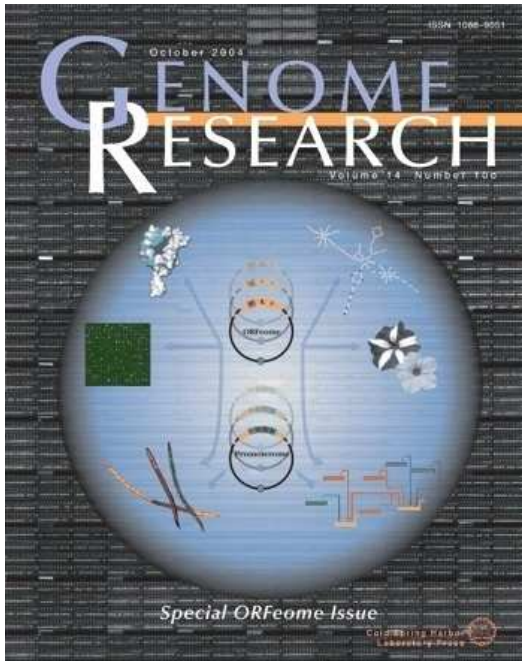
Chi-Hao Luan,[1,3] Shihong Qiu,[1] James B. Finley,[1] Mike Carson,[1] Rita J. Gray,[1] Wenying Huang,[1] David Johnson,[1] Jun Tsao,[1] Jérôme Reboul,[2] Philippe Vaglio,[2] David E. Hill,[2] Marc Vidal,[2] Lawrence J. DeLucas,[1] and Ming Luo[1,3]

[1]*Center for Biophysical Sciences and Engineering, Southeast Collaboratory for Structural Genomics, University of Alabama at Birmingham, Birmingham, Alabama 35294, USA;* [2]*Center for Cancer Systems Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, and Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA*

Proteome-scale studies of protein three-dimensional structures should provide valuable information for both investigating basic biology and developing therapeutics. Critical for these endeavors is the expression of recombinant proteins. We selected *Caenorhabditis elegans* as our model organism in a structural proteomics initiative because of the high quality of its genome sequence and the availability of its ORFeome, protein-encoding open reading frames (ORFs), in a flexible recombinational cloning format. We developed a robotic pipeline for recombinant protein expression, applying the Gateway cloning/expression technology and utilizing a stepwise automation strategy on an integrated robotic platform. Using the pipeline, we have carried out heterologous protein expression experiments on 10,167 ORFs of *C. elegans*. With one expression vector and one *Escherichia coli* strain, protein expression was observed for 4854 ORFs, and 1536 were soluble. Bioinformatics analysis of the data indicates that protein hydrophobicity is a key determining factor for an ORF to yield a soluble expression product. This protein expression effort has investigated the largest number of genes in any organism to date. The pipeline described here is applicable to high-throughput expression of recombinant proteins for other species, both prokaryotic and eukaryotic, provided that ORFeome resources become available.

14:2102–2110 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

→ Pour Interactome



14:2201–2206 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

Resource

Generation of the *Brucella melitensis* ORFeome Version 1.1

Amélie Dricot,[1] Jean-François Rual,[1,2] Philippe Lamesch,[1,2] Nicolas Bertin,[2] Denis Dupuy,[2] Tong Hao,[2] Christophe Lambert,[1] Régis Hallez,[1] Jean-Marc Delroisse,[1] Jean Vandenhaute,[1] Ignacio Lopez-Goñi,[3] Ignacio Moriyon,[3] Juan M. Garcia-Lobo,[4] Félix J. Sangari,[4] Alastair P. MacMillan,[5] Sally J. Cutler,[5] Adrian M. Whatmore,[5] Stephanie Bozak,[6] Reynaldo Sequerra,[6] Lynn Doucette-Stamm,[6] Marc Vidal,[2] David E. Hill,[2] Jean-Jacques Letesson,[1] and Xavier De Bolle[1,7]

[1]Research Unit in Molecular Biology (URBM), University of Namur, 5000 Namur, Belgium; [2]Center for Cancer Systems Biology and Department of Cancer Biology, Dana-Farber Cancer Institute and Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA; [3]Departamento de Microbiologia, Universidad de Navarra, Pamplona 31008, Spain; [4]Departamento de Biología Molecular, Universidad de Cantabria, Santander 39011, Spain; [5]Department of Statutory and Exotic Bacterial Diseases, FAO/WHO Collaborating Centre for Reference and Research on Brucellosis, Veterinary Laboratories Agency, Weybridge, Surrey KT15 3NB, United Kingdom; [6]Agencourt Biosciences Corporation, Beverly, Massachusetts 01915, USA

→ Pour Interactome

Gateway & Plant biology

Technologie gateway & RNAi

ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Methods 30 (2003) 289–295

METHODS

www.elsevier.com/locate/ymeth

Constructs and methods for high-throughput gene silencing in plants

Chris Helliwell and Peter Waterhouse*

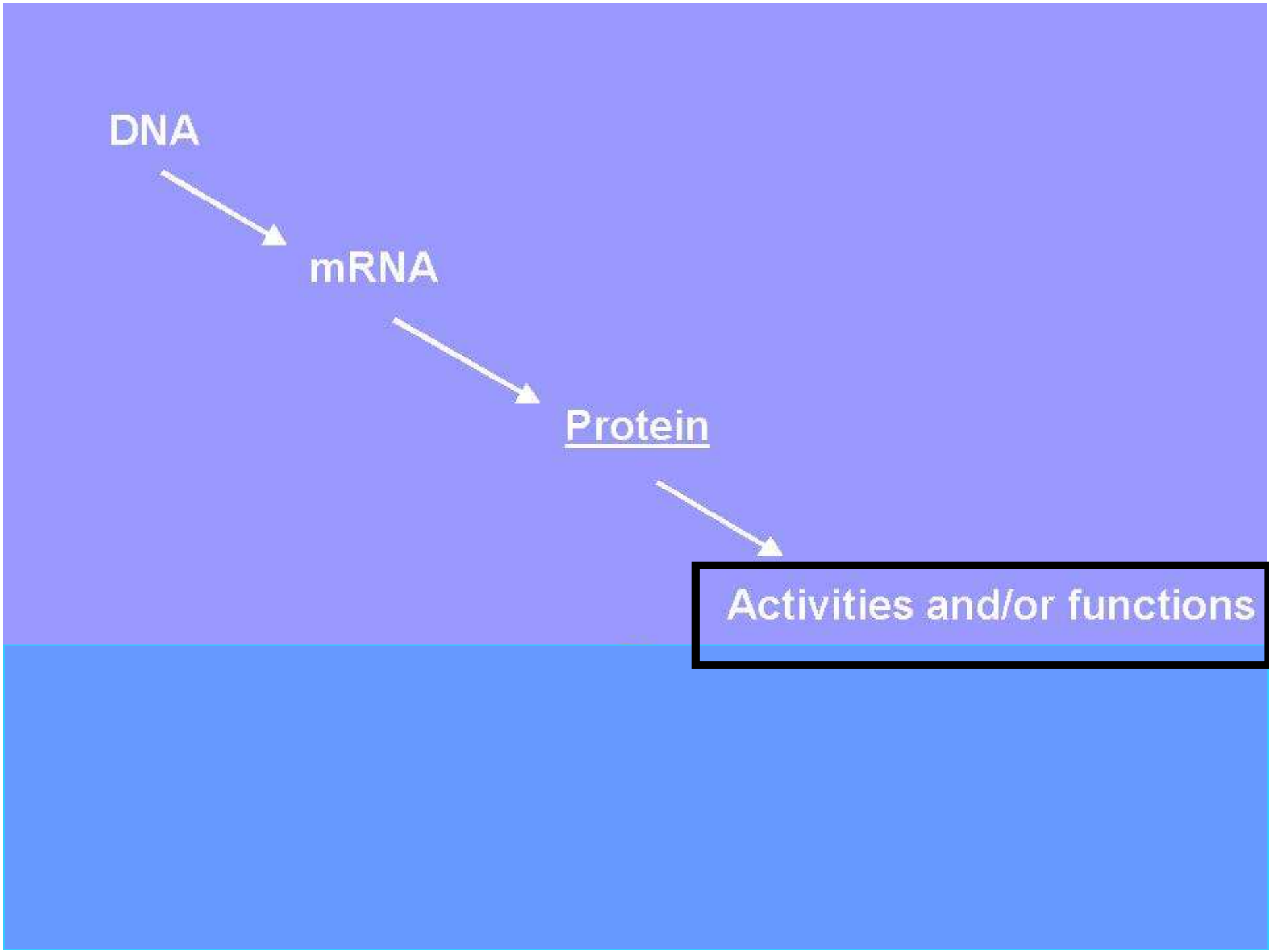
CSIRO Plant Industry, GPO Box 1600, Canberra ACT 2601, Australia

Accepted 7 February 2003

Abstract

Gene silencing can be achieved by transformation of plants with constructs that express self-complementary (termed hairpin) RNA containing sequences homologous to the target genes. The DNA sequences encoding the self-complementary regions of hairpin (hp) RNA constructs form an inverted repeat. The inverted repeat can be stabilized in bacteria through separation of the self-complementary regions by a "spacer" region. When the spacer sequence encodes an intron, the efficiency of gene silencing is very high. There are at least three ways in which hpRNA constructs can be made. The construct may be generated from standard binary plant transformation vectors in which the hairpin-encoding region is generated de novo for each gene. Alternatively, generic gene-silencing vectors such as the pHANNIBAL and the pHELLSGATE series can be used. They simply require the insertion of PCR products, derived from the target gene, into the vectors by conventional cloning or by using the Gateway directed recombination system. In this article, we describe and evaluate the advantages of these vectors and then provide the protocols for their efficient use.

Keywords: Gene silencing; Recombination cloning; Plant functional genomics; pHELLSGATE


DNA
mRNA
Protein
Activities and/or functions

Annu. Rev. Plant Biol. 2003. 54:669–89
doi: 10.1146/annurev.arplant.54.031902.135014

METABOLOMICS IN SYSTEMS BIOLOGY

Wolfram Weckwerth
Max-Planck-Institut für Molekulare Pflanzenphysiologie, 14424 Potsdam, Germany; email: weckwerth@mpimp-golm.mpg.de

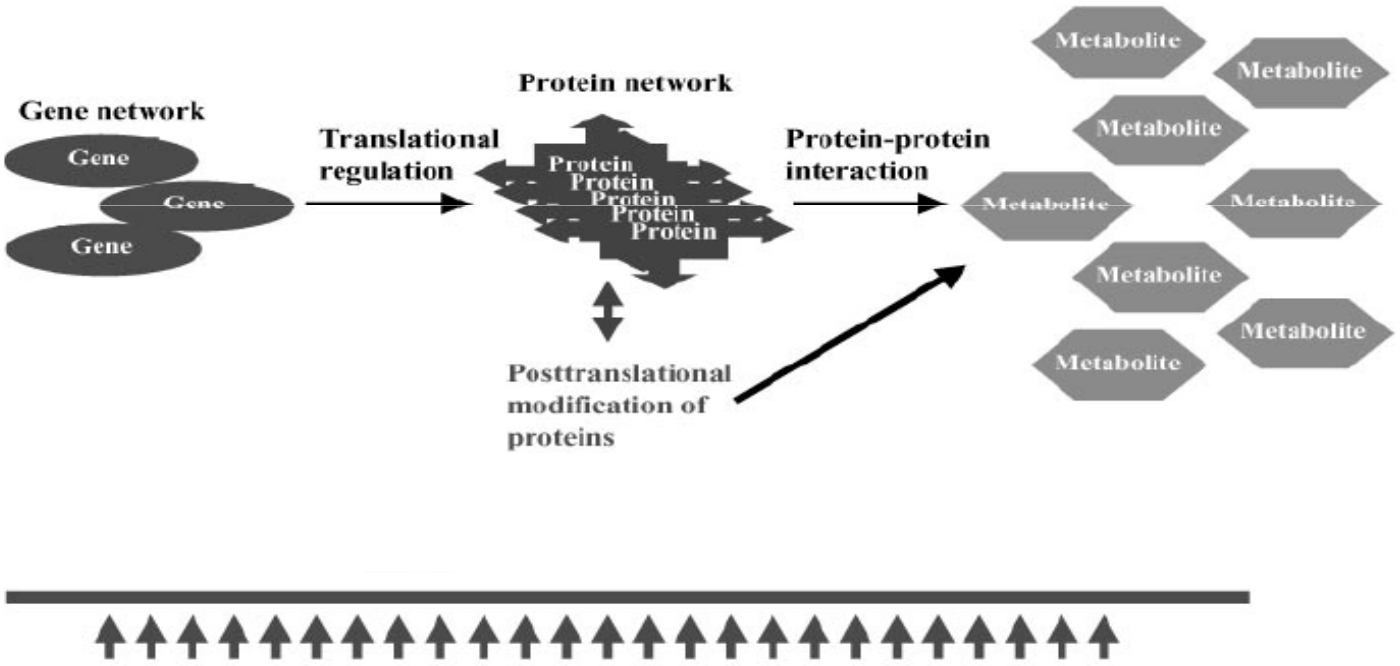


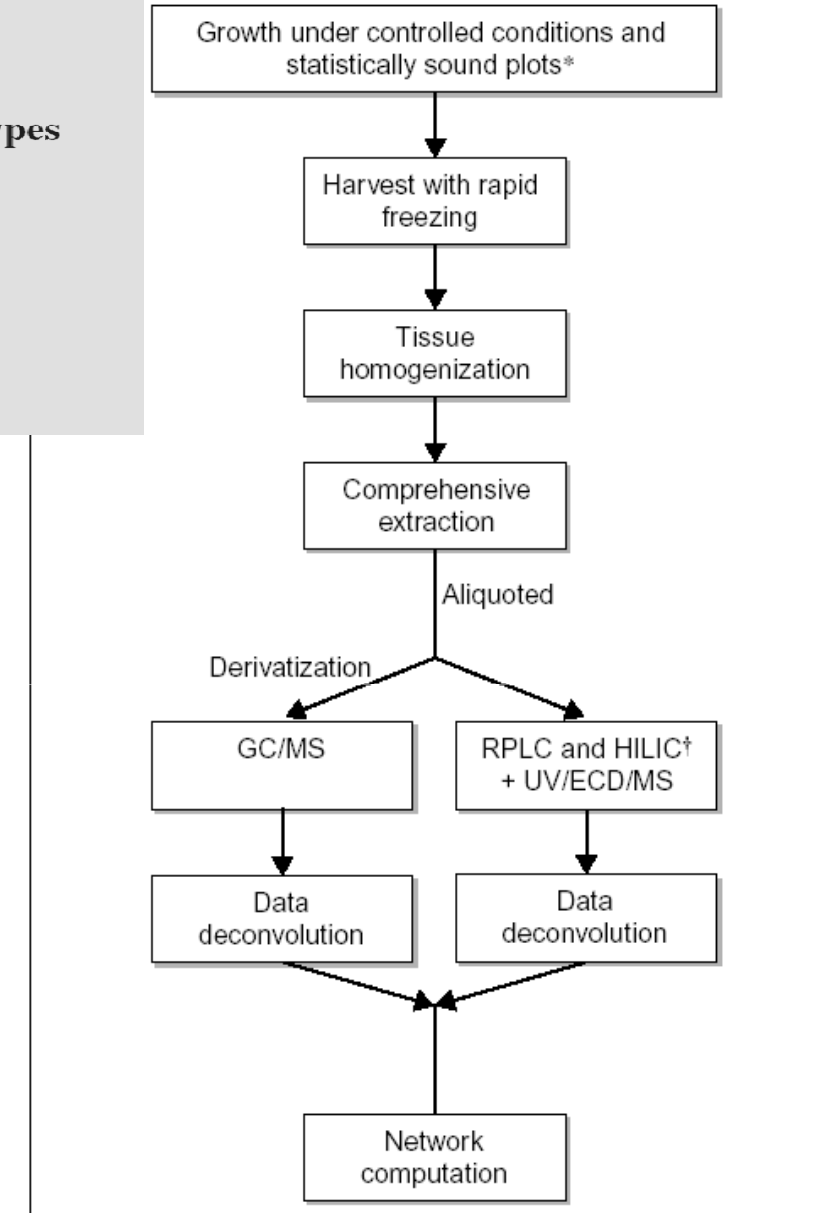
Figure 1 Amplification of a metabolic network and feedback regulation in response to developmental and environmental conditions.

Plant Molecular Biology **48**: 155–171, 2002.

Metabolomics – the link between genotypes and phenotypes

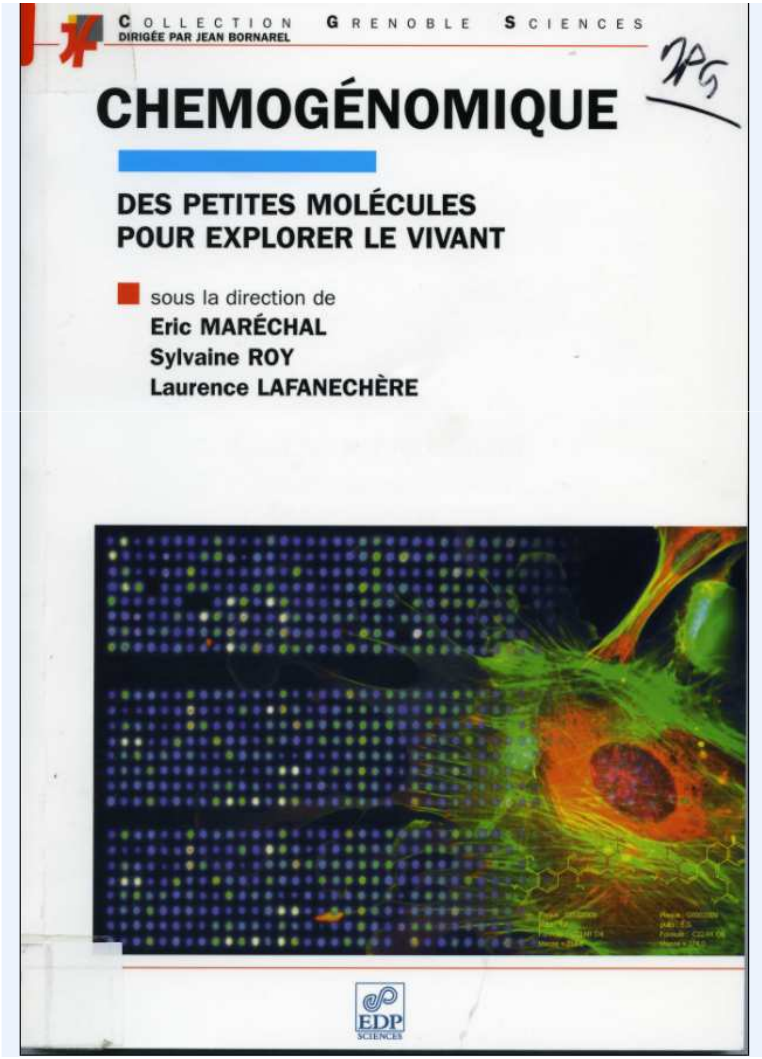
Oliver Fiehn
Max Planck Institute of Molecular Plant Physiology, 14424 Potsdam, Germany
(e-mail fiehn@mpimp-golm.mpg.de)

Key words: functional genomics, mass spectrometry, metabolism, metabolite profiling



Proposed scheme for comprehensive metabolomic data acquisition.

La chémogénomique



= criblage à haut débit de molécules naturelles ou de synthèses pour découvrir de nouvelles cibles thérapeutiques et de nouveaux médicaments

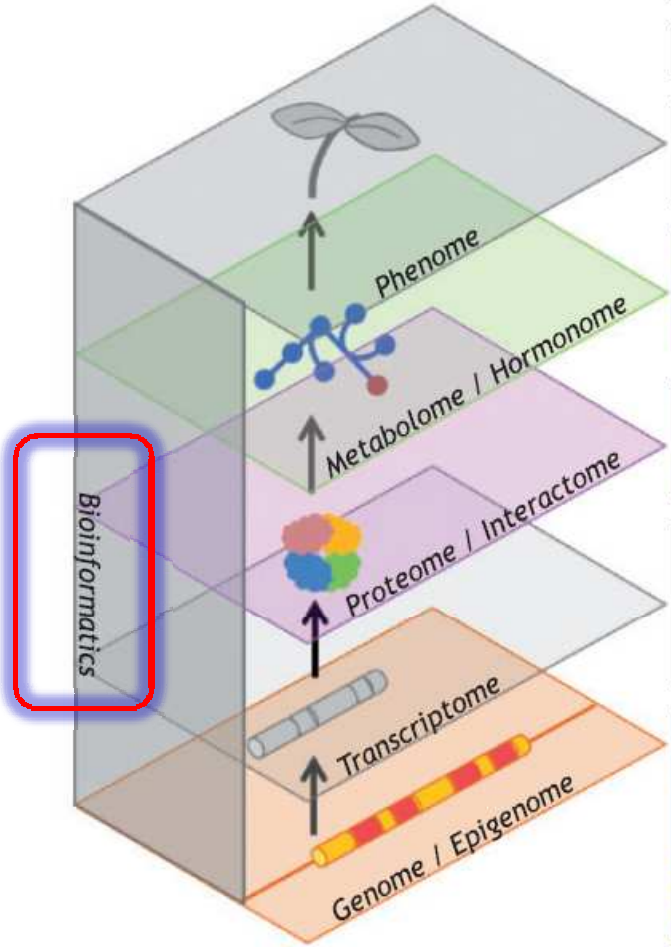
…

Ou de nouvelles molécules interférant avec Fonctions cellulaires ou des molécules d'intérêt en agronomie et microbiologie (pesticides, herbicides, etc …)

- Les chimiothèques

- les criblages de molécules et la mesure de l'effet biologique

- Intérêt en post-génomique



Omics instances	Resources
Integrated database	TAIR[1]
Mutant lines	FOX line[2], Ac/Ds tag line[3], T-DNA tag line[4], TILLING[5]
Natural variations	NASC[6], ABRC[7]
Metabolic map	PMN AraCyc[8], Reactome[9]
Metabolome profiles	PRIMe[10], Golm Metabolome Database[11]
Hormonome profile	RiceFOX DB[2]
Proteome / modificome profiles	RIPP-DB[12], PPDB[13], PhosPhAt[14]
Subcellular localization	PODB2[15], SUBAII[16], NASC proteome database[17]
Interactome maps	TAIR[1], AtPID[18], Plant Interactome Database[19]
Full-length cDNA clones, ESTs	RAFL clones[20], RARGE[21]
Expression profiles	AtGenExpress[22], Genevestigator[23]
Non-coding RNA profiles	Arabidopsis MPSS[24]
Co-expression network	ATTEDII[25]
Genome sequence, gene annotation	TAIR[1]
Re-sequencing	Arabidopsis 1001 genome project[26]
Focused gene family database (eg. Transcription factor)	RARTF[27], AGRIS[28], DATF[29]
DNA methylome	SIGnAL[30]
Chromatin epigenome	EPIC web site[31]

Fig. 1 An updated omic space with emerging omics layers: epigenome, interactome and hormonome added to each of the closely related layers with illustrative resources for *Arabidopsis* available on the web. [1]http://www.arabidopsis.org/, [2]http://ricefox.psc.riken.jp/, [3]http://rarge.gsc.riken

K. Mochida and K. Shinozaki
Plant Cell Physiol. 52(12): 2017–2038 (2011)