

Gerstein Bioinformatics Group Memb	
File Edit View Go Communicator Help	
[MB&B] YALE UNIVERSITY Gerstein Lab	
Group Member	Title
Mark Gerstein	Principal I
Suganthi Balasubramanian	Postdoctoral
Paul Harrison	Postdoctoral
Hedi Hegyi	Postdoctoral
Jochen Junker	Postdoctoral
Yuval Kluger	Sloan Fell
Ning Lan	Postdoctoral
Nicholas Luscombe	Postdoctoral
Jiang Qian	Postdoctoral
Vadim Alexandrov	Graduate
Paul Bertone	Graduate
(Joint with Michael Snyder)	
Rajdeep Das	Graduate
Dov Greenbaum	Graduate
Ronald Jansen	Graduate
Ted C Johnson	Graduate
Werner G Krebs	Graduate
Nathaniel Echols	
Jimmy Lin	
Patrick McGarvey	
Christopher Titcombe	Undergra
Joann Delvecchio	Administ

MB&B Bioinformatics Group (Gerstein Lab)

Gerstein Group - Yale Bioinformatics - Netscape

File Edit View Go Communicator Help

Bookmarks Location <http://bioinfo.mbb.yale.edu/people> What's Related

[MB&B] YALE UNIVERSITY

Gerstein Lab BIOINFORMATICS

About us, Address, CSB
 GeneCensus Comparative Genomics
 Macromolecular Movements Database
 Lectures, Papers
 People, Jobs available
 Search, Fun stuff

We do research in the emerging field of bioinformatics, using computation to analyze genome sequences and macromolecular structures. [\[More on PI\]](#)

Research on Genomes & Structures

Comparative genomics is the first focus of the lab. Much of this work is encapsulated in our [GeneCensus](#) system, which presents the results from various analyses comparing the genomes sequenced so far.

Our genomics work rests on the utilization of a library of [aligned cores](#) for the ~400 currently known protein folds, the basic "building blocks" in nature.

Macromolecular motions is our second focus. To further analyze the variability of each of the building blocks, we have developed a [database system](#) for collecting motions and measuring structural variability.

Our motions analysis includes web [movies](#) generated by a server and free software for analyzing macromolecular [geometry](#), particularly packing.

Other Information

We disseminate our research through published [papers](#) and on-line [lectures](#) and courses.

Search Only this site (best) [Special Search](#)

Document: Done

Who we are?

Scientific American: Feature Article: The Bioinformatics Gold Rush: July...

File Edit View Go Communicator Help

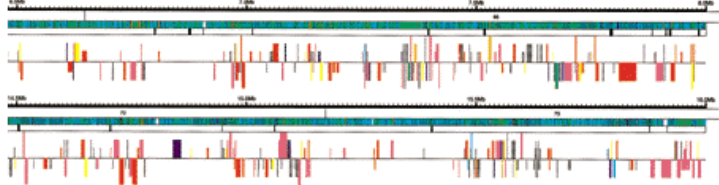
Bookmarks Netsite: 0700issue/0700howard.html What's Related

SCIENTIFIC AMERICAN

Main Menu Interview Bookmarks Feedback
 Current Issue Explore! Ask the Experts Marketplace Search the

FEATURE ARTICLES

The Bioinformatics Gold Rush



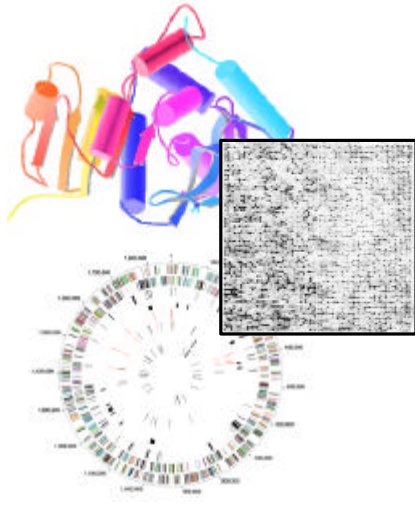
A \$300-million industry has emerged around turning raw genome data into knowledge for making new drugs

By Ken Howard

SUBTOPICS: Plastics." When a family friend whispered this word to Dustin Hoffman's character in the 1967 film *The*

Document: Done

Seems to becoming of interest...



What do we do?



(Molecular) **Bio - informatics** =

Conceptualizing **biology in terms of molecules** (in the physical-chemical sense) and then **applying “informatics”** techniques (derived from disciplines such as CS and statistics) to **organize and analyze** the **information** associated with these molecules, on a **large-scale** and in an **integrative fashion**. Bioinformatics is a practical discipline with many **applications**.

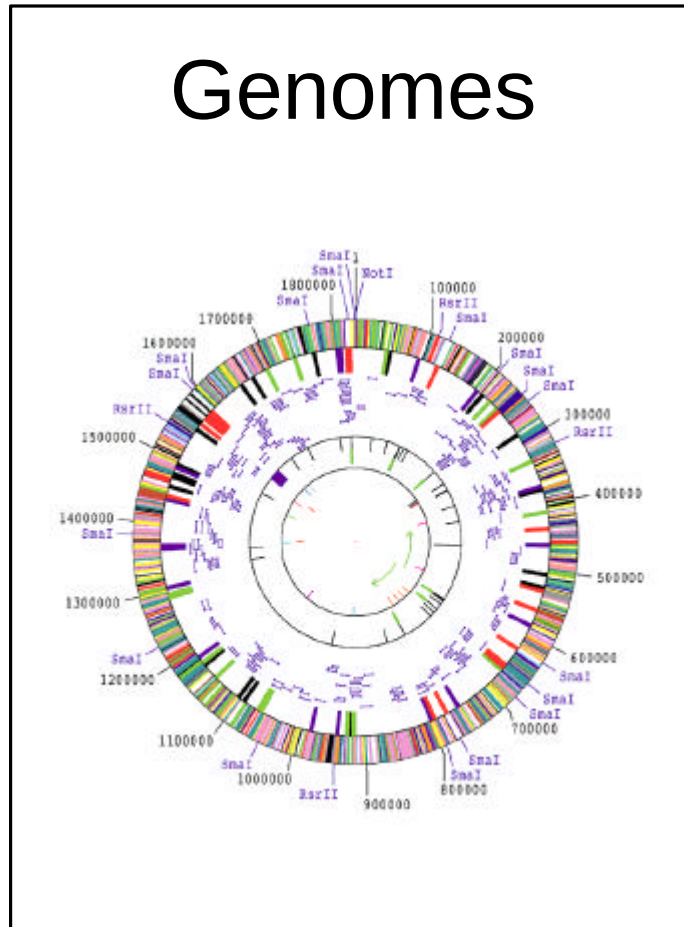
Large-scale, Integrative Analysis of...

Structures



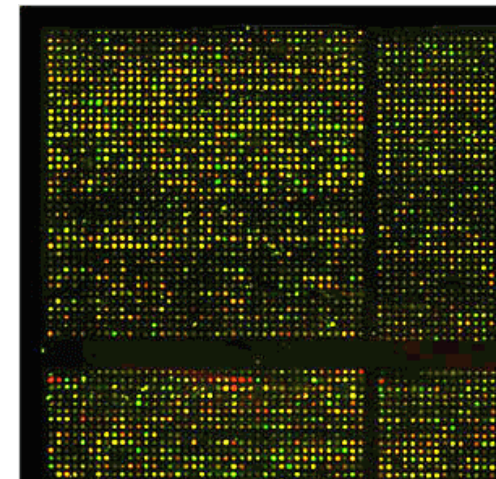
1990

Genomes



2000

Expression Data



2010

1995

Bacteria,
1.6 Mb,
~1600 genes
[*Science* **269**: 496]



1997

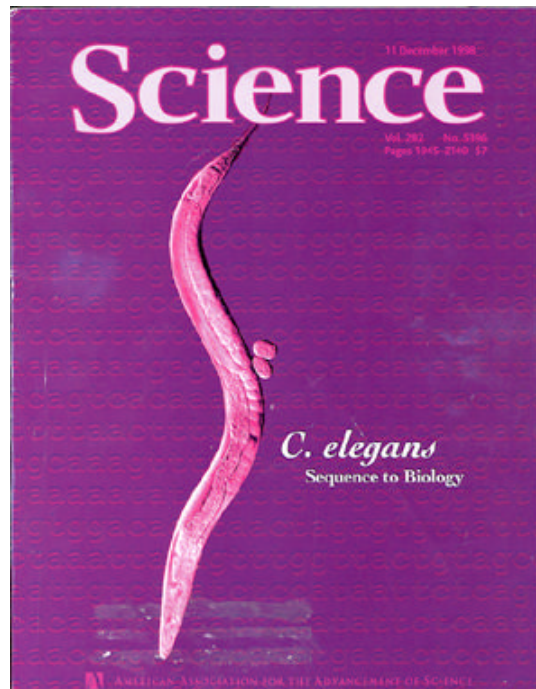
Eukaryote,
13 Mb,
~6K genes
[*Nature* **387**: 1]



2000
The Year
of
Genomics

1998

Animal,
~100 Mb,
~20K genes
[*Science* **282**:
1945]



2000

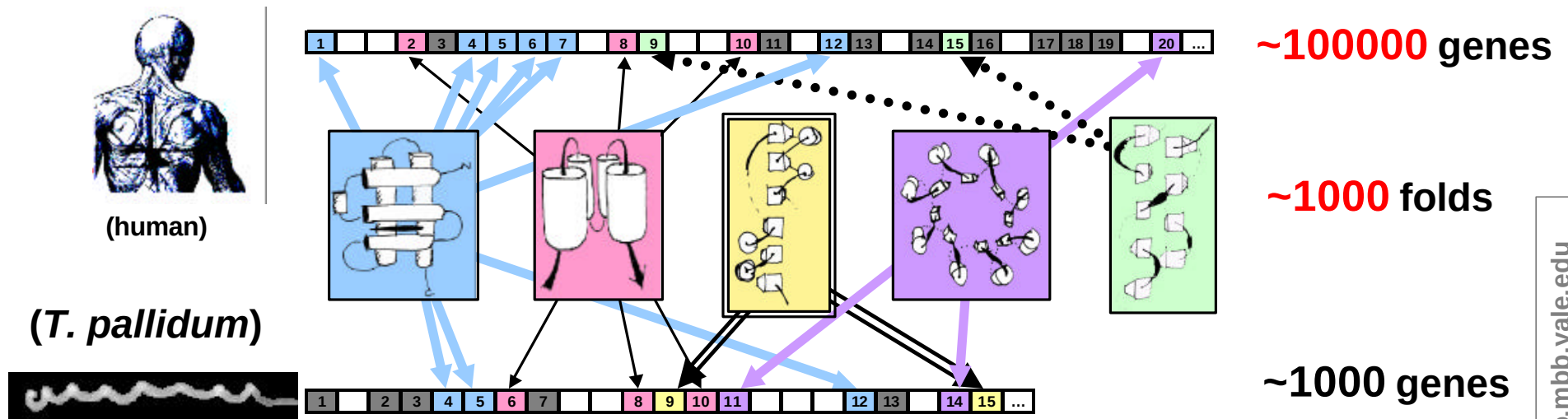
Human,
~3 Gb,
~100K
genes [???



'98 spoof

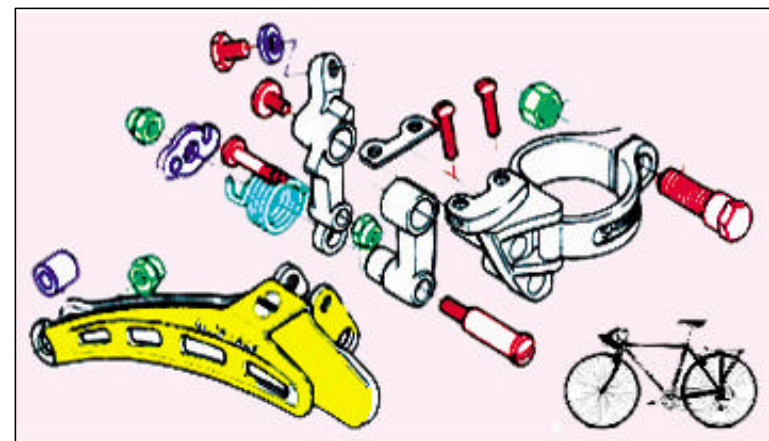
real thing, Apr '00

The Finite Parts List as a Simplifying Theme

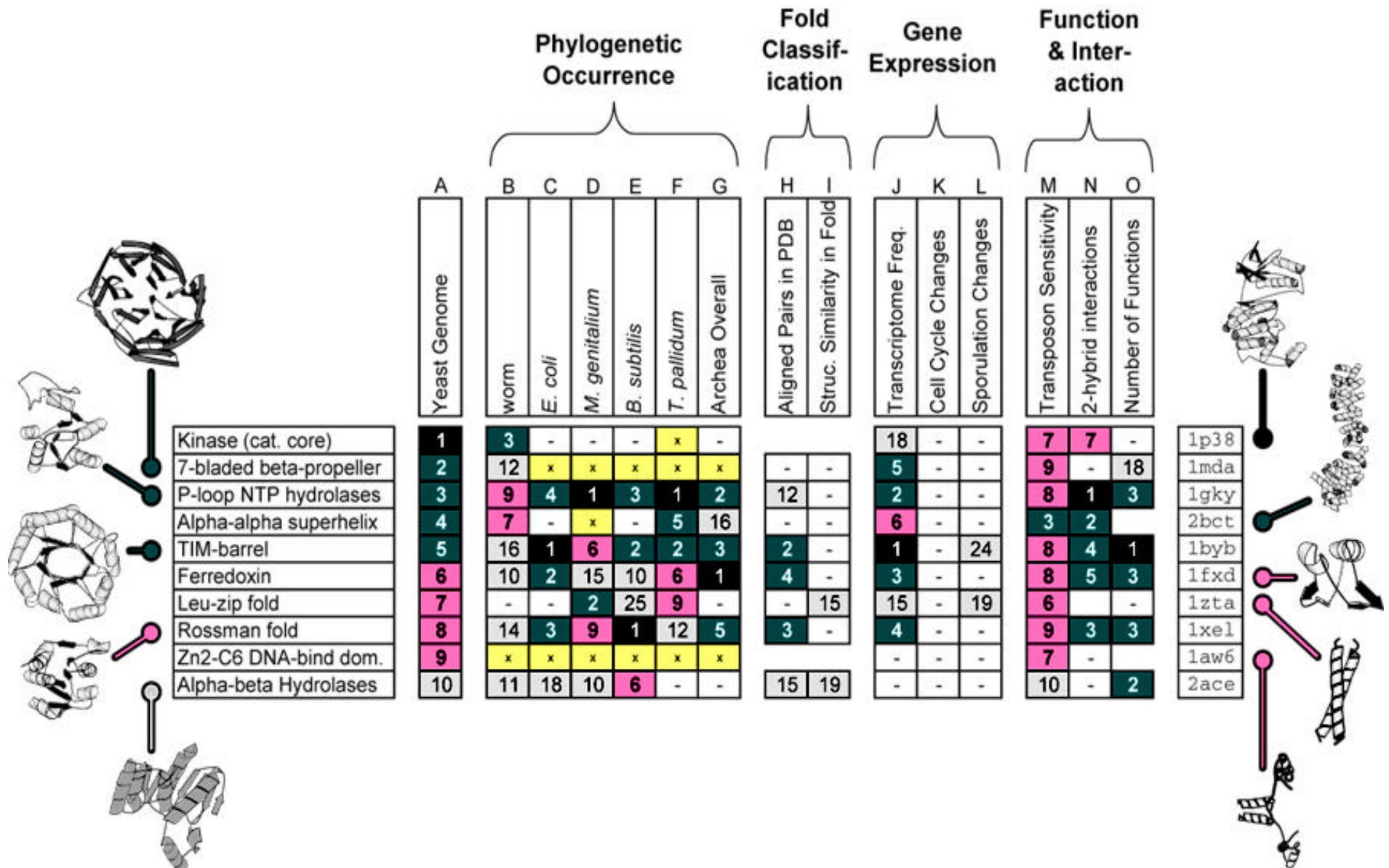


Parts = **FOLDS**, pathways, functions, sequence families, blocks, motifs....

Global Surveys of a
Finite Set of Parts from
Many Perspectives



Surveying a Finite Parts List from a Infinite Number of Perspectives



PartsList Ranking Viewers

RANKINGS 30 Highest Values -- cele -- occurrences - Netscape

YALE GERSTEIN LAB

30 Highest Values -- Fold Occurrences

• [Caenorhabditis elegans](#)

30 Highest Ranked

Out of 248	Representative Domain	20 Genomes	Abu	Mia	Mha	Pho	Ser	Cal	Axa	Syn	Eso	Bak	Mub	Hnf	Hov	Mam	Mns	Rbu
1	d1ajv	881	2	2	2	0	4	829	2	11	18	2	2	0	2	0	0	2
2	d1dec	576	0	0	0	3	3	558	1	1	0	2	1	2	0	1	1	2
3	d3lck	636	2	3	2	3	129	454	1	13	1	3	15	0	1	1	1	0
4	d1tsg	323	0	0	0	0	1	322	0	0	0	0	0	0	0	0	0	0
5	d1tfo	297	0	0	0	1	10	284	0	0	1	0	0	0	0	0	1	0
6	d2lbd	257	0	0	0	0	0	257	0	0	0	0	0	0	0	0	0	0
7	d1a17	494	0	21	7	4	114	250	25	42	6	1	0	1	0	0	0	6
8	d1asp2	321	1	2	0	2	77	238	0	0	0	0	0	1	0	0	0	0
9	d1dai	1301	56	62	61	61	130	223	57	68	73	64	54	51	46	34	36	69
10	d2aw0	1426	93	87	583	21	114	207	31	38	83	27	44	34	21	3	3	6
11	d1cvi	420	13	1	2	5	39	175	3	24	19	33	70	5	1	4	5	2
12	d1qotb	323	0	0	0	0	140	168	0	10	0	0	1	1	0	0	0	0
13	d1tcp	138	0	0	0	0	0	138	0	0	0	0	0	0	0	0	0	0
14	d1eny	785	31	16	26	19	72	132	33	59	79	98	126	25	21	4	5	5
15	d1a51_1	315	6	10	6	7	31	130	4	15	24	23	29	13	4	1	1	1
16	d1aj2	858	45	38	43	30	84	123	28	51	93	88	77	35	22	7	12	20
16	d1bor	152	0	0	0	0	29	123	0	0	0	0	0	0	0	0	0	0
18	d1ldr	117	0	0	0	0	0	116	0	0	0	0	0	0	0	0	0	0
19	d1rec	129	0	0	0	0	14	112	0	1	0	0	0	0	2	0	0	0
20	d1pty	122	0	0	0	0	10	111	0	0	1	0	0	0	0	0	0	0
21	d1mek	289	7	2	4	4	33	110	13	22	22	17	13	14	6	1	1	1

Rankings2 - Netscape

YALE GERSTEIN LAB

search

Rankings²

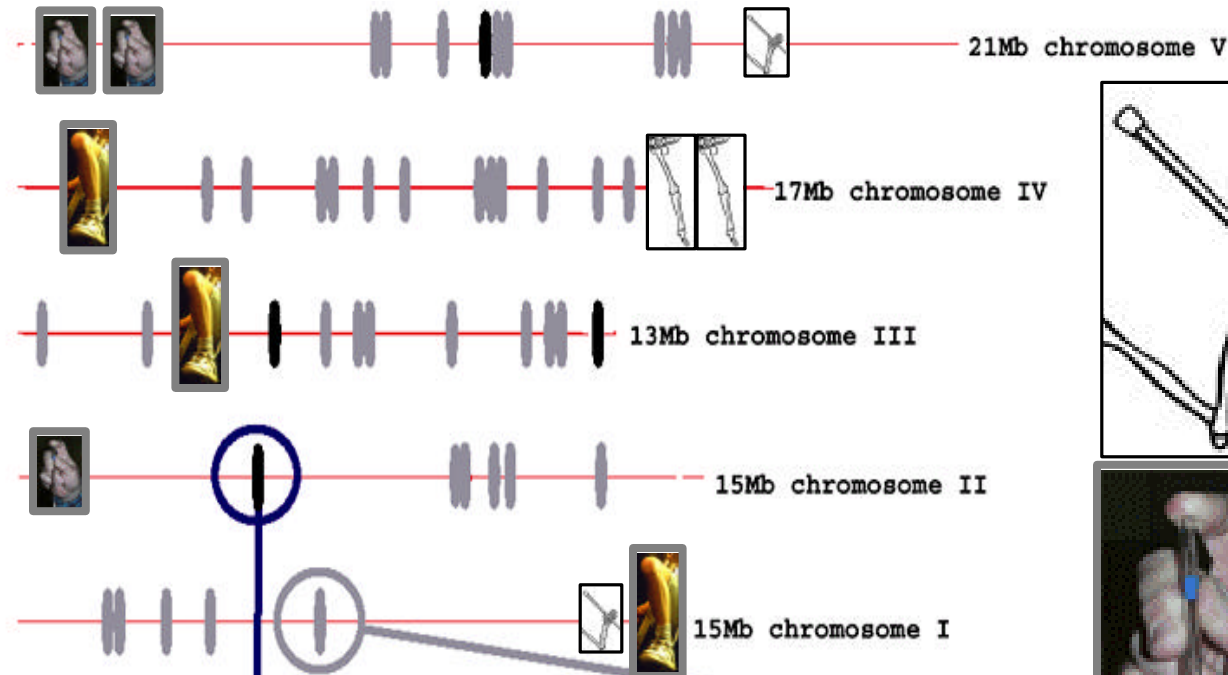
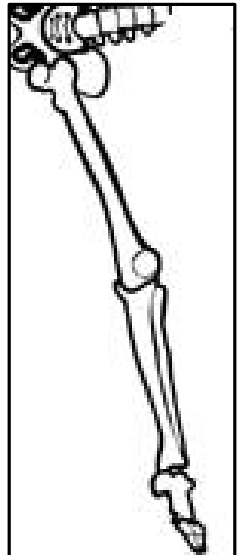
	fold occurrence in Scer	fold occurrence in Cele	fold percentage in reference genome	fold percentage in reference transcriptome
Max of all	140	829	11.3	11.1
Min of all	0	0	0.1	0
Average	6.9	22.9	0.5	0.5
Non zero hits	215	247	213	134
Rank again				
d1ajw :1.002.001 Immunoglobulin-like beta-sandwich Class: All beta proteins	4	829	0.2	0.2
d1dec :1.007.003 Knottins (Small inhibitors) Class: Small Proteins	3	556		
d3lck :1.005.001 Protein kinases (PK) Class: Multi-domain (alpha and beta) proteins	129	454	6.4	0.9
d1tsg :1.004.105 C-type lectin-like Class: Alpha plus beta proteins	1	322		
d1tfo :1.007.033 Glucocorticoid receptor-like (DNA-binding domain) Class: Small Proteins	10	284	0.4	0.1
d2lbd :1.001.095 Ligand-binding domain of nuclear receptor Class: All alpha proteins	0	257		
d1a17 :1.001.091 alpha-alpha superhelix Class: All alpha proteins	114	250	4.3	2.3
d1asp2 :1.007.031 Classic zinc finger Class: Small Proteins	77	238	1.4	0.2

J Qian,
B Stenger,
J Lin....



Rank Folds by Genome
Occurrence, Expression, Fold
Clustering, Length, &c

Pseudogenomics: Surveying “Dead” Parts

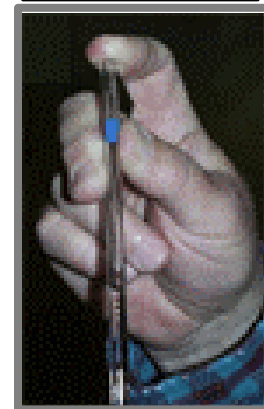
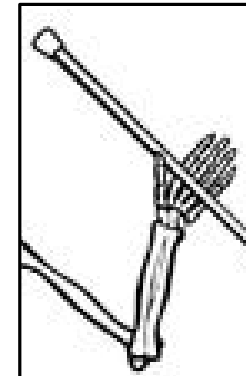


pseudogene fragment on worm chromosome II

```
TKRTSNGFGQDVVVDLFSILDSGLVARAHXVLQDIFEFFAS
KKMVTIFS#APHSPHSAPHYCAQFDNSAATVKV
```

a paralog with the homologous segment highlighted (from chromosome I)
(W09C3.6, serine/threonine protein phosphatase PP1)

```
MTAPMDVDNLMSRLNLNVGMSGGRLTTSVNEQELQTCCAVAKSVFASQASLLEVEPPIVC
GDIHGQYSDLLRIFDKNGFPDVFNLFLG DYVDRGRQNIETICLMLCFKIKYPENFFMLR
GNHECPAINRVYGFYEENRRYKSTRLWSIFQDTFNWMPCLGLIGSRILCMHGGLSPHLQ
TLDOLROLPRPODPPNPSIGIDLLWADPDOWVKGWOANTRGVS YVFGODVVADVCSRLDI
DLVARAHOVVODGYEFFASKMVTIFSAPHYCGOFDNSAATMKVDENMVCTFVVMYKPTPK
SMRRG*
```



Example of
a potential
 Ψ G with
frameshift in
mid-domain