Correlation of Microarray Probes give Evidence for Mycoplasma Contamination in Human Studies

W. B. Langdon

CREST

Department of Computer Science

MedGEC 2013

RN/12/11

26.6.2013

Mycoplasma Contamination in Public Databases

- Background:
  - BioTechniques 2009 article
  - Mycoplasma
  - Affymetrix microarray
  - USA government GEO database

GEO
Gene Expression Omnibus

- Evidence:
  - Human microarray probes which match mycoplasma
  - Correlation between probes across GEO
- Implications for EC researchers

Mycoplasma genes in the Human Genome

Summary

- Mycoplasma contaminate human sample
- DNA, including Mycoplasma DNA, is sequenced
- Mar 2000 Mycoplasma gene added to GenBank labelled “homo sapiens unknown sequence”
- April 2001 unknown EST sequence added by Affymetrix to HG-U133 +2 microarray
- 2008 Mycoplasma contamination of 2 of 3 replicants leads to 1570561_at being differentially expressed.
- Suspicion about “unknown human EST” leads to BioTechniques article (Dec 2009)

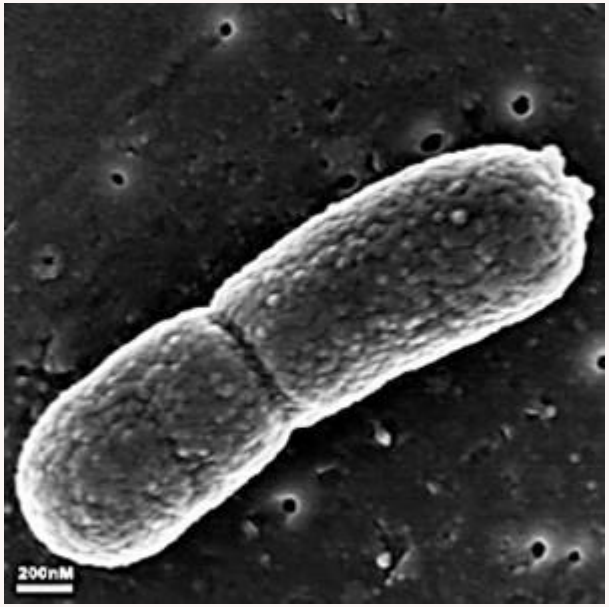
History
Affymetrix HG-U133 plus 2 probeset 1570561_at

- BioTechniques 2009 article showed Affymetrix human microarray probeset 1570561_at measures expression of a Mycoplasma gene not human gene.
- ≈1% of published data in GEO came from samples contaminated with mycoplasma.
- **Other probes also show mycoplasma**

Mycoplasma

- Tiny bacteria which routinely infect microbiology laboratories
- Not easy to detect
- Mycoplasma infection makes sample measurements useless
- Mycoplasma infects 10-25% laboratory cultures.
- 30+ mycoplasma genomes have been sequenced

mycoplasma capricolum



Affymetrix HG-U133 +2

- First single microarray to measure expression of all human genes
- Short DNA strands on chip are designed to be complementary to expressed gene which stick to them.
- Stuck DNA fluoresces and hence chip can be read by laser.
- 11um feature size, noisy, so:
- Typically 11 measurements (probes) per DNA sequence

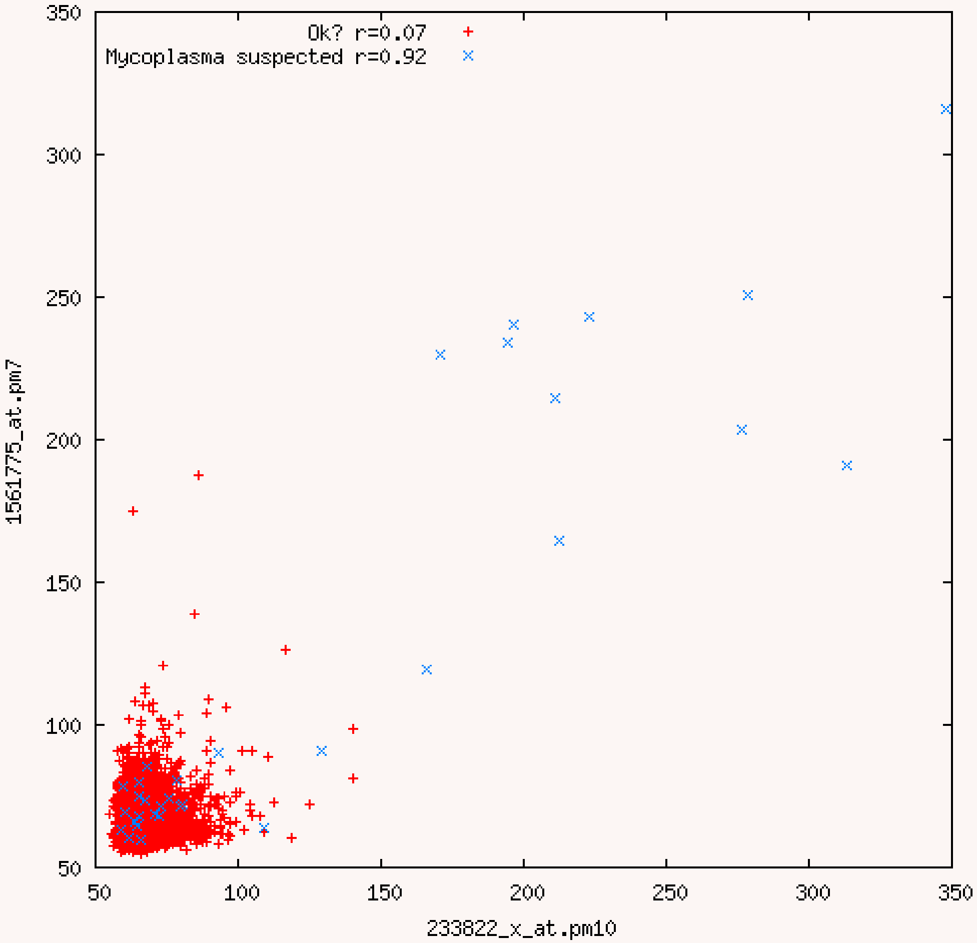
Gene Expression Omnibus

- US government’s GEO is an archive containing ≈1 million gene expression datasets.
- All HG-133 +2 datasets were loaded into RNAnet
- RNAnet allows instant access to normalised microarray data

Mapping probes to Mycoplasma

- DNA sequences on HG-U133 +2 known
- They are intended to align to human genome.
- Bowtie used to try to align all 1208516 probes against all mycoplasma genomes
- 437 match, but consider only 106 exact matches
- Restrict to 61 with strong signal in GEO

GEO scatter plot 2 probes matching mycoplasma

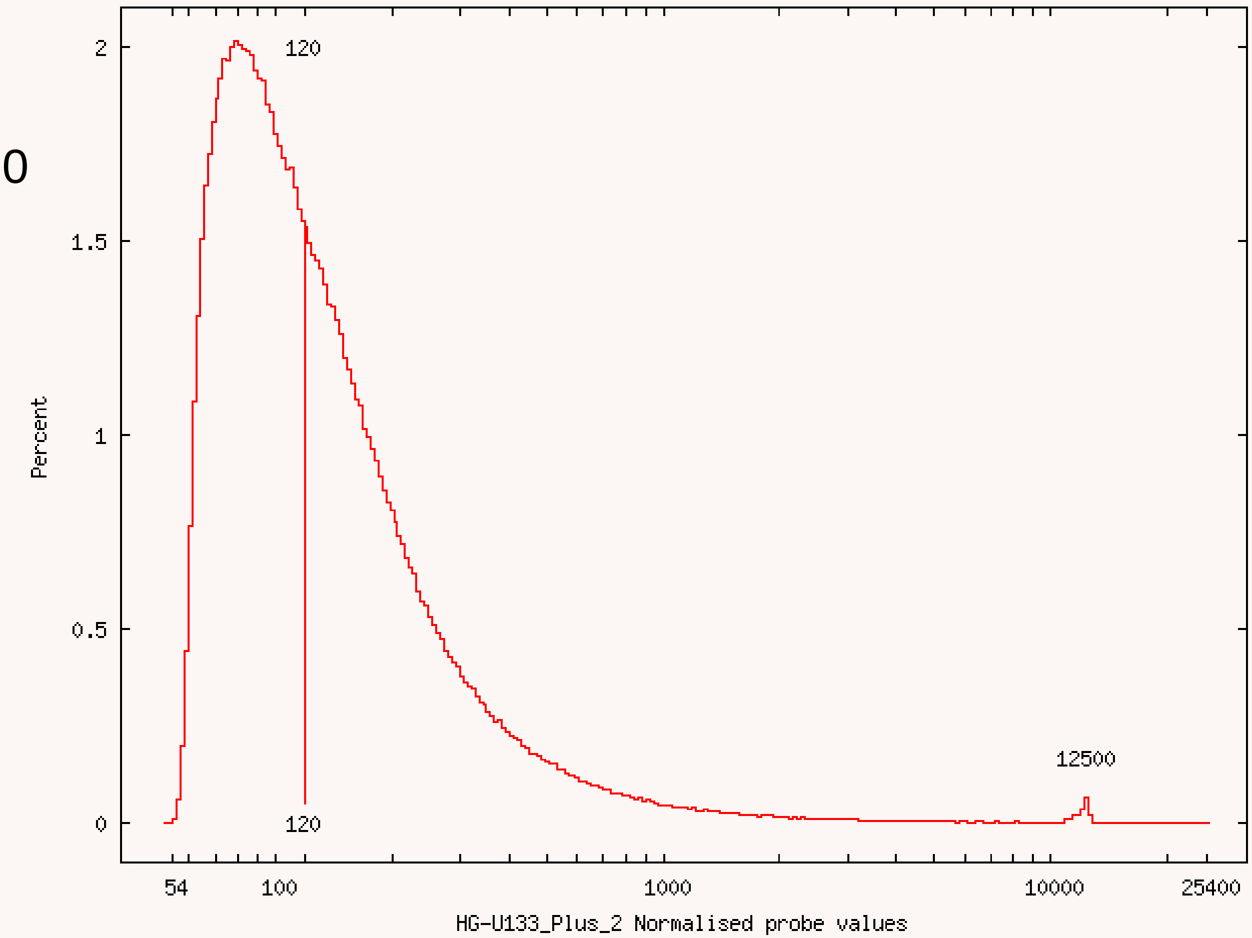


33 samples suspected of contamination R=0.92

Remaining 2724 R=0.07

Normalised HG-U133 +2 probes

54% are <120
Median 112
Mode 79

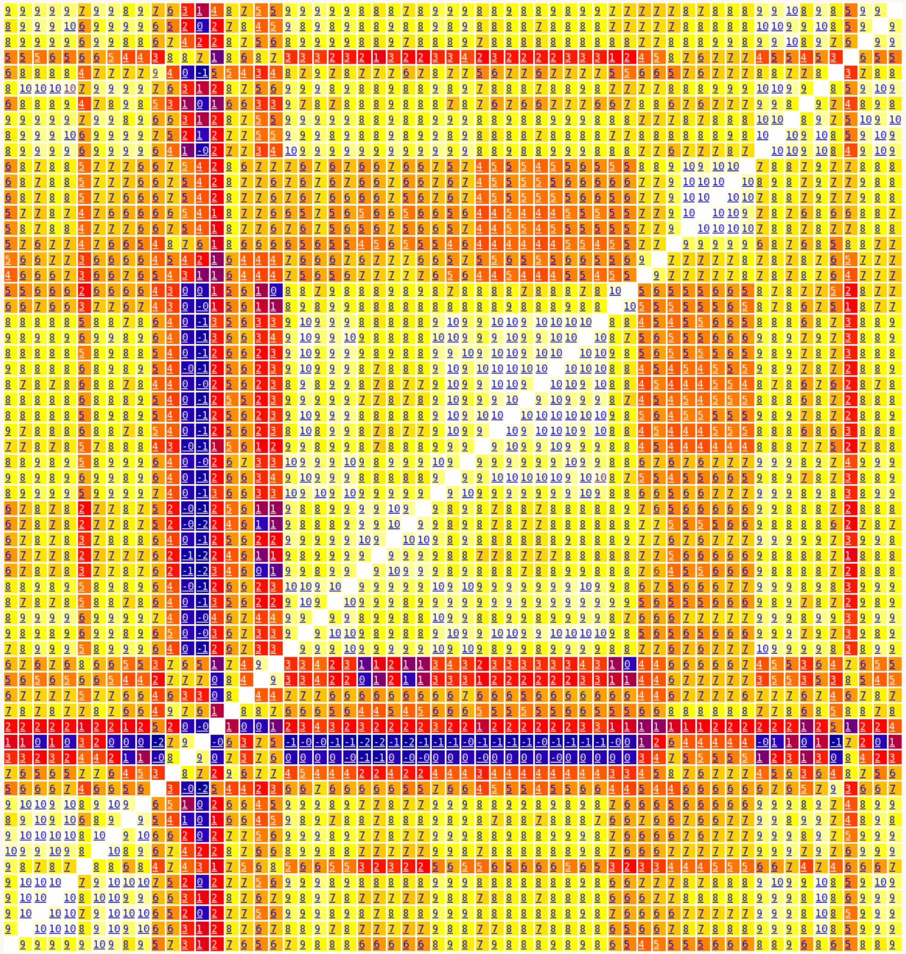


Correlation of 61 probes

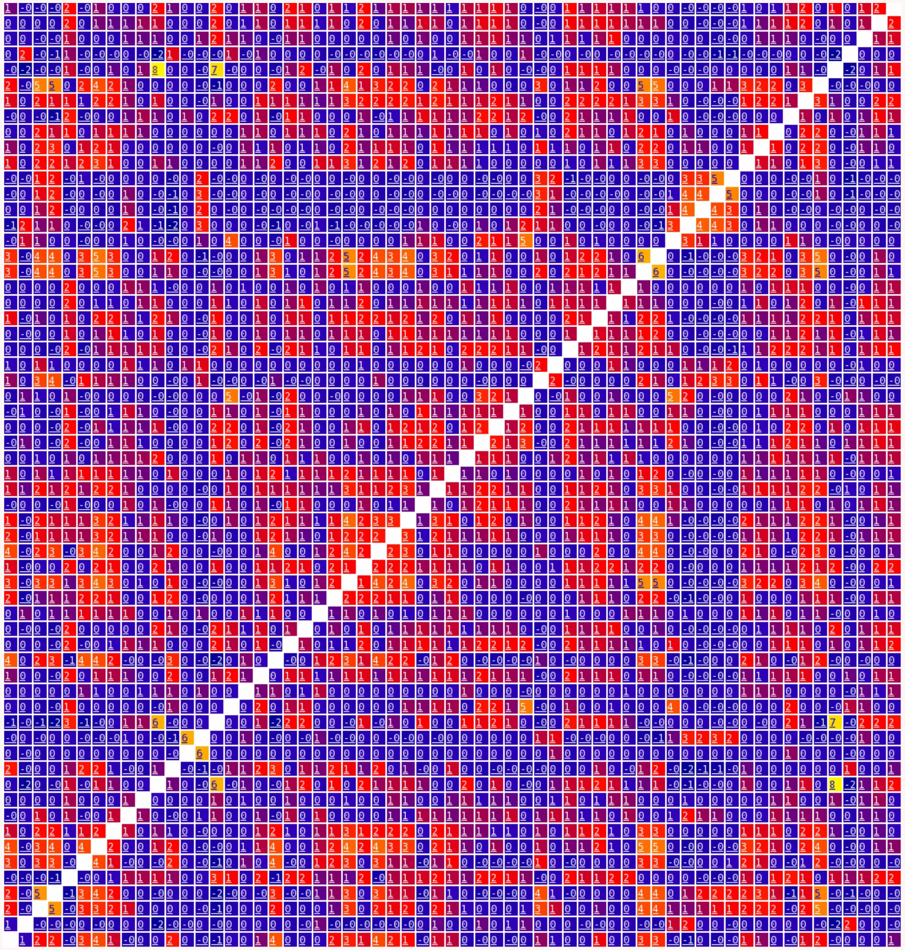
- Correlation in GEO of 61 probes with each other. 1830 pairs.
- In 0.7% mycoplasma contaminated *all* pairs are correlated.

HG-U133 +2.0 correlation Mycoplasma suspected in GEO

Essentially all 61 probes are correlated with all the other 60.

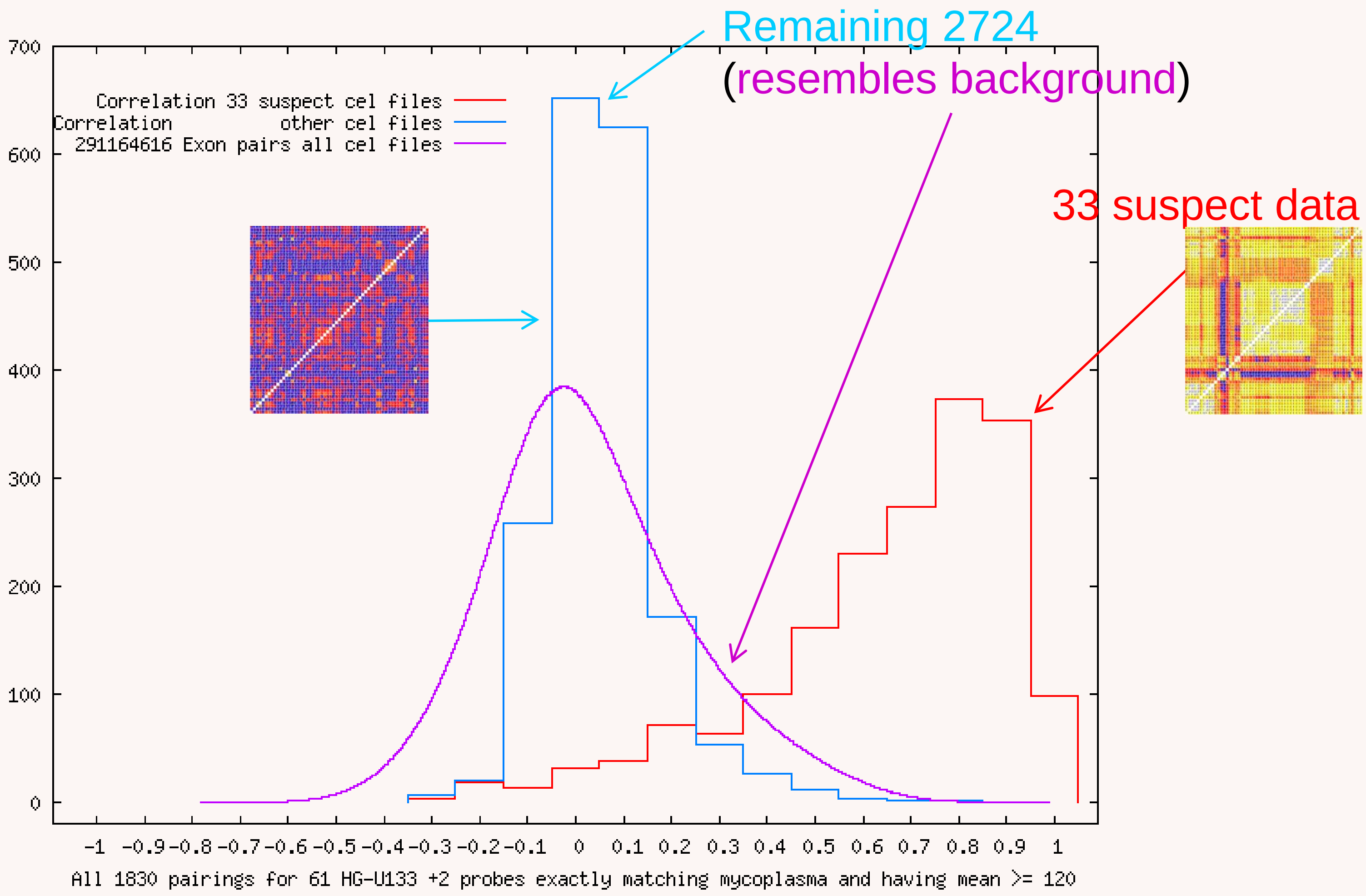


HG-U133 +2.0 correlation rest of GEO



Essentially no correlation

HG-U133 +2 mycoplasma probes



Implications for MedGEC

- Medicine increasingly reliant on computer algorithms and databases.
- Exponential growth in public bioinformatics databases
- Creators of data may not pass knowledge of data’s quality to curators or users.
- Biologist say computer scientists must use “Due diligence”.

➢ Cannot take most important bioinformatics database on trust

Summary

- Computer scientists must use “due diligence” with public bioinformatics data.
- Mycoplasma disrupts human gene expression. 26-100% contamination in articles
- All HG-U133 +2 probes which map to at least one mycoplasma genome and are expressed are highly correlated in suspect GEO data.
- Probably due to Mycoplasma signal dominating that of human genes.

END

http://www.cs.ucl.ac.uk/staff/W.Langdon/ http://www.epsrc.ac.uk/ EPSRC

GO annotations are from Affymetrix’ netaffy and so assume human genes. Mycoplasma hyorhinis HUB-1 gene ids from NCBI NC 014448.1

Probe-set	GO biological process term	GO molecular function term	Symbol	HUB-1 description
224354_at	glucose metabolic process, oxidation reduction	glyceraldehyde-3 phosphate dehydrogenase (phosphorylating) activity protein binding NAD or NADH binding	gap	Glyceraldehyde 3-phosphate dehydrogenase C
1567703_at			rpmF	50S ribosomal protein L32
233847_x_at			ribF	Riboflavin biosynthesis protein
234623_x_at			*As 234432_at*	
234432_at			MHR_0358	hypothetical protein
1561775_at			MHR_0246	hypothetical protein
233822_x_at	tRNA aminoacylation for protein translation	nucleotide binding aminoacyl-tRNA ligase activity ATP binding	serS	Seryl-trna synthetase protein
1570561_at	first reported mycoplasma probeset		16S-23S ribosomal RNA intergenic spacer.	
211690_at	rRNA processing translational elongation TOR signaling cascade ribosomal small subunit biogenesis glucose homeostasis positive regulation of apoptosis	structural constituent of ribosome protein binding	MHR_r0001	16S ribosomal RNA
1555623_at	oxidation reduction	oxidoreductase activity FAD or FADH2 binding	MHR_0008	dihydrolipoamide dehydrogenase

Growing number of DNA sequences

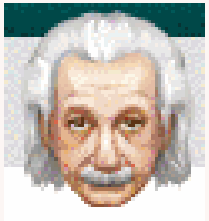
- The number of sequences is growing exponentially.
  - “Moore’s Law” no. of DNA bases in GenBank doubles approximately every 18 months
  - 24,656 taxa already sequenced RefSeq,2013
- Known problem. Nobody working on a solution? Will only get worse.
- Contamination in other direction Human genes → other species
- Many human genes in non-primate DNA sequence databases

Mycoplasma Genes in the Human Genome

- “Unexpected presence of mycoplasma probes on human microarrays”, BioTechniques, Dec 2009
- 2nd example “More Mouldy Data: Virtual Infection of the Human Genome”, technical report RN/11/14.
- Multiple human genes in other (non-human) organisms’ DNA sequence databases

Technical Report RN/11/14 Virtual Infection of the Human Genome

- arXiv blog, blogspot, Slashdot

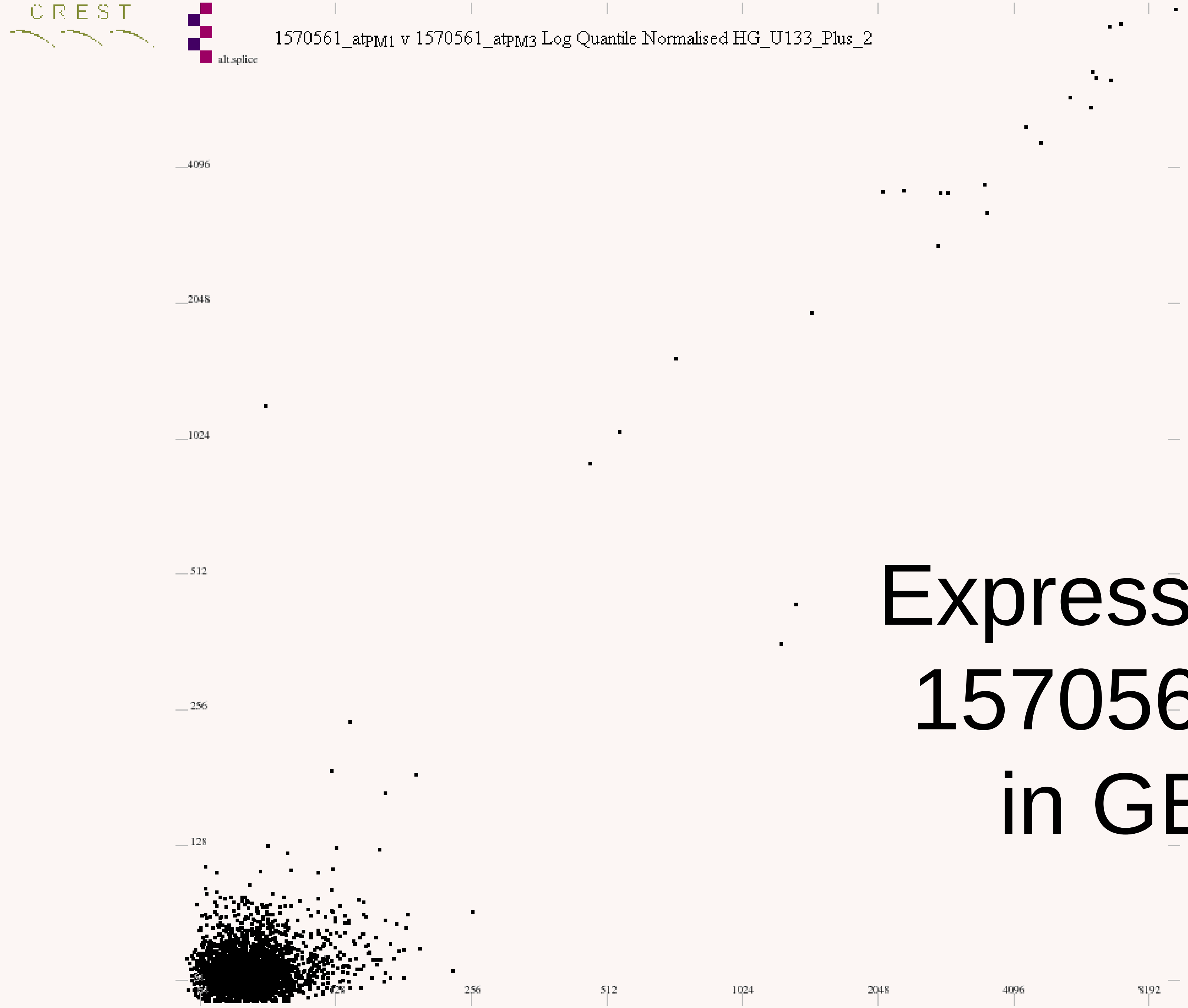


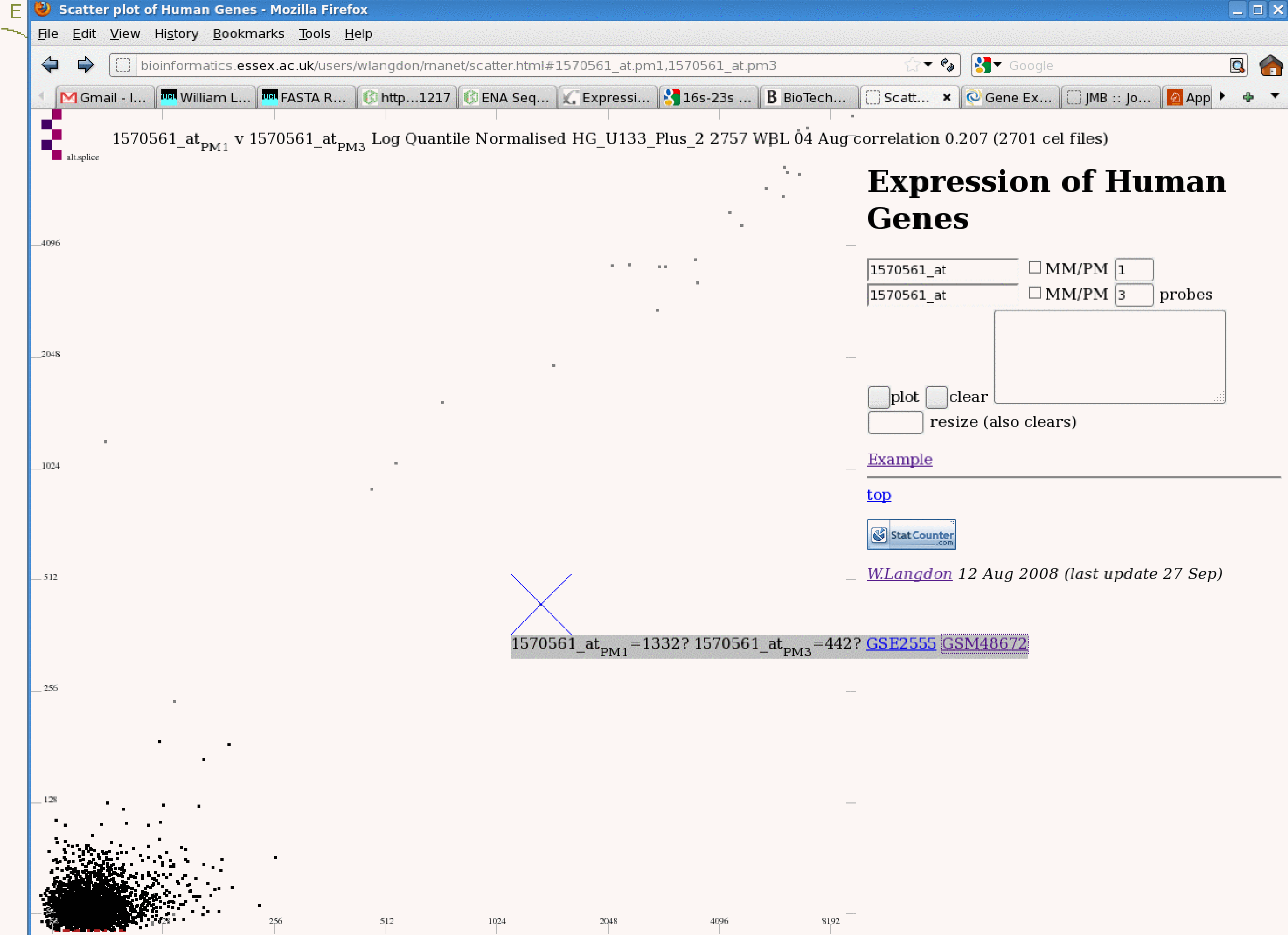
- SCIENCE ORF.at KURZWEIL ACCELERATING INTELLIGENCE Blog
- Der Spiegel, 4 July, New Scientist 13 July

Expression of 1570561_at in GEO

- RNAnet http://bioinformatics.essex.ac.uk/users/wlangdon/rnanet/scatter.html#1570561_at.pm1,1570561_at.pm3
- To show values across 2757 samples plot two probes (of 11) against each other.
- 31 of 33 high expression values come from cell cultures (94% v. 34% back ground).

Expression of 1570561_at in GEO





1570561_at$_{PM1}$ v 1570561_at$_{PM3}$ Log Quantile Normalised HG_U133_Plus_2 2757 WBL 04 Aug correlation 0.207 (2701 cel files)

Expression of Human Genes

1570561_at □ MM/PM 1

1570561_at □ MM/PM 3 probes

plot clear

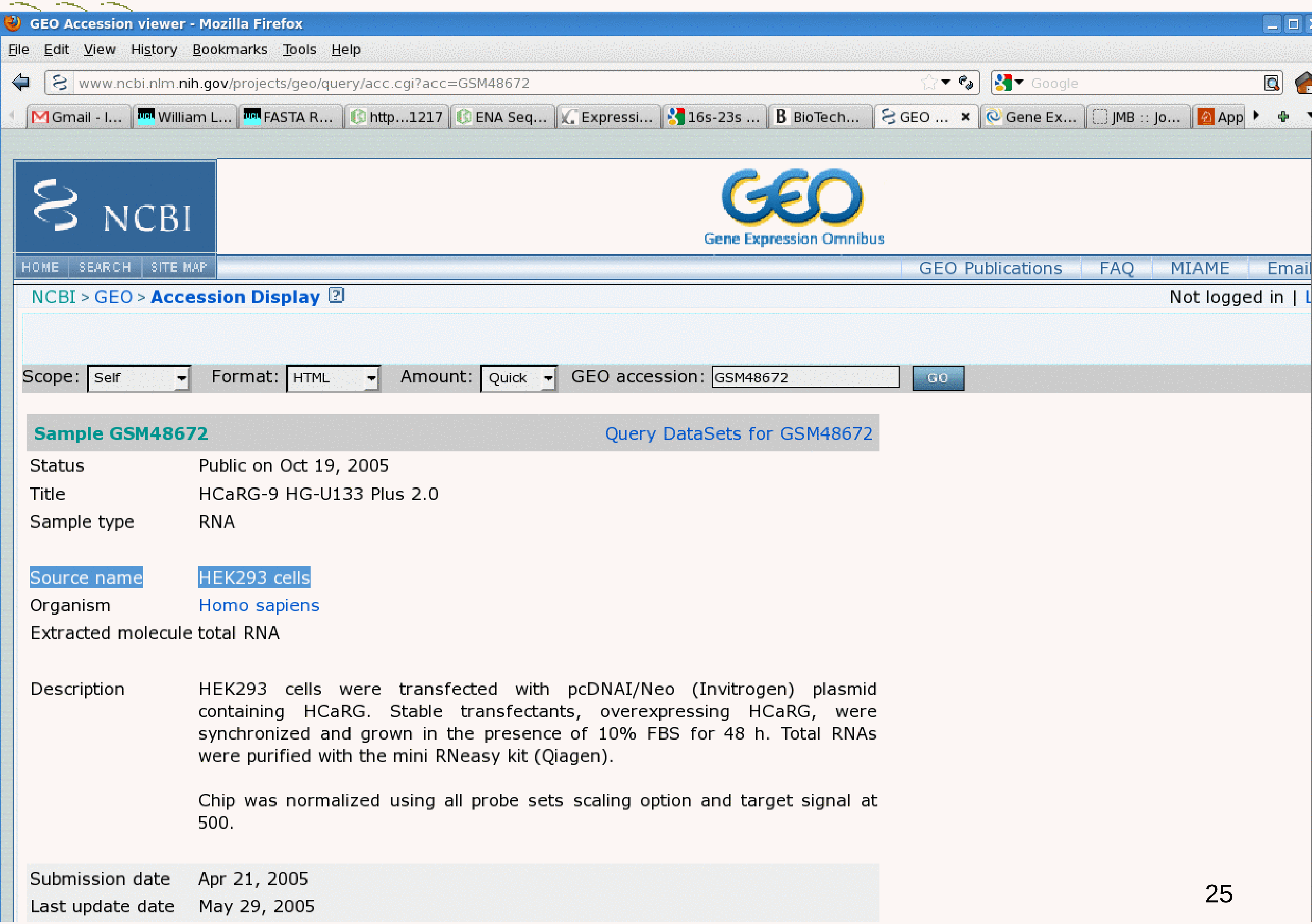
resize (also clears)

Example

top

W.Langdon 12 Aug 2008 (last update 27 Sep)

1570561_at$_{PM1}$=1332? 1570561_at$_{PM3}$=442? GSE2555 GSM48672


GEO Accession viewer - Mozilla Firefox
File Edit View History Bookmarks Tools Help
www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSM48672
Google
Gmail - I... William L... FASTA R... http...1217 ENA Seq... Expressi... 16s-23s ... BioTech... GEO ... Gene Ex... JMB :: Jo... App
NCBI
GEO
Gene Expression Omnibus
HOME SEARCH SITE MAP
GEO Publications FAQ MIAME Emai
NCBI > GEO > Accession Display
Not logged in | L
GO
Query DataSets for GSM48672

Scope: Self Format: HTML Amount: Quick GEO accession: GSM48672

Sample GSM48672

Status	Public on Oct 19, 2005
Title	HCaRG-9 HG-U133 Plus 2.0
Sample type	RNA
Source name	HEK293 cells
Organism	Homo sapiens
Extracted molecule	total RNA
Description	HEK293 cells were transfected with pcDNAI/Neo (Invitrogen) plasmid containing HCaRG. Stable transfectants, overexpressing HCaRG, were synchronized and grown in the presence of 10% FBS for 48 h. Total RNAs were purified with the mini RNeasy kit (Qiagen). Chip was normalized using all probe sets scaling option and target signal at 500.
Submission date	Apr 21, 2005
Last update date	May 29, 2005

Another Mycoplasma in GenBank?

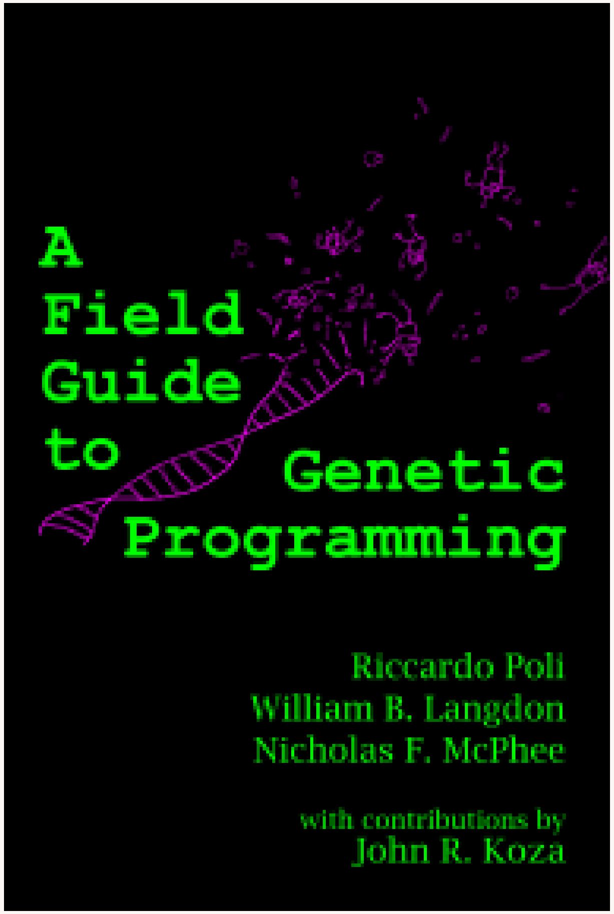
- 2011 AF241217 Blast run again
  - GenBank has not fixed error
  - All match Mycoplasma except 1st and 34th DA466599
- Second example: DA466599
  - DA466599 matches various species of Mycoplasma
  - DA466599 uploaded into Data Bank of Japan 2 years after HG-U133 +2 was launched
- DA466599 also Mycoplasma 16S-23S ribosomal RNA intergenic spacer labelled as Human in GenBank

Genes Spread

- Microbes infect microbiology laboratories
- 2 genes have been copied into GeneBank
  - 1 via Japan, 1 into commercial tool. Others? patents?
  - Many human genes in nonprimate databases
- Data are routinely copied, allowing virtual genes (venes) to spread globally.
- Laboratories routinely sterilise glassware. They do not sterilise their databases.

A Field Guide To Genetic Programming

http://www.gp-field-guide.org.uk/



A Field Guide to Genetic Programming

Riccardo Poli
William B. Langdon
Nicholas F. McPhee

with contributions by
John R. Koza

Free PDF

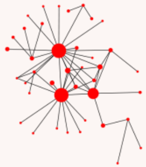
Free E-book

The Genetic Programming Bibliography

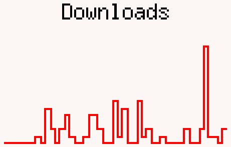
http://www.cs.bham.ac.uk/~wbl/biblio/

8755 references and 8351 online publications

RSS Support available through the Collection of CS Bibliographies. XML RSS

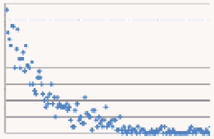

Part of gp-bibliography 84 40 Revision:1.1794 29 May 2011

A web form for adding your entries.
Co-authorship community. Downloads



A personalised list of every author’s GP publications.

blog.html



Search the GP Bibliography at
http://liinwww.ira.uka.de/bibliography/Ai/genetic.programming.html