In Silico Infection of the Human Genome

W. B. Langdon

CREST

Department of Computer Science

EvoBio 2012, pp245-249

Non Human Genes in GenBank Public Database of the Human Genome

- Background: BioTechniques article
  - Mycoplasma
  - Affymetrix microarray
  - NCBI databases

NCBI GenBank

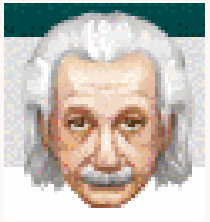
- Evidence:
  - Blast DNA sequence comparisons
  - Gene expression levels in GEO via RNAnet
- Implications

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

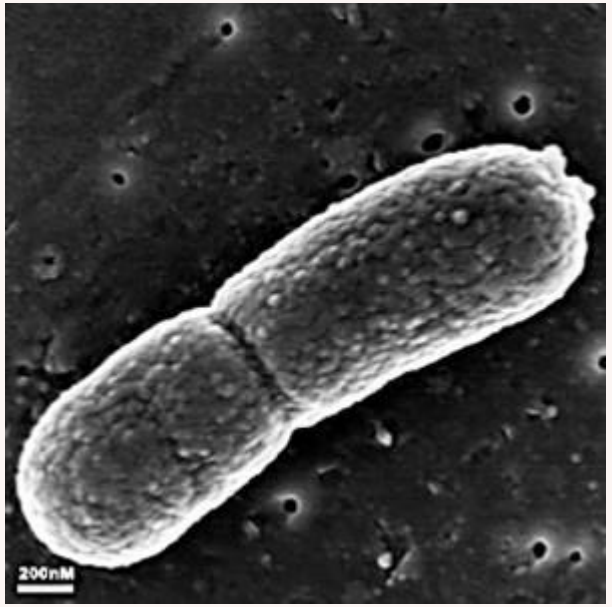
- arXiv blog, blogspot, Slashdot
- SCIENCE ORF.at KURZWEIL ACCELERATING INTELLIGENCE Blog
- Der Spiegel, 4 July, New Scientist 13 July



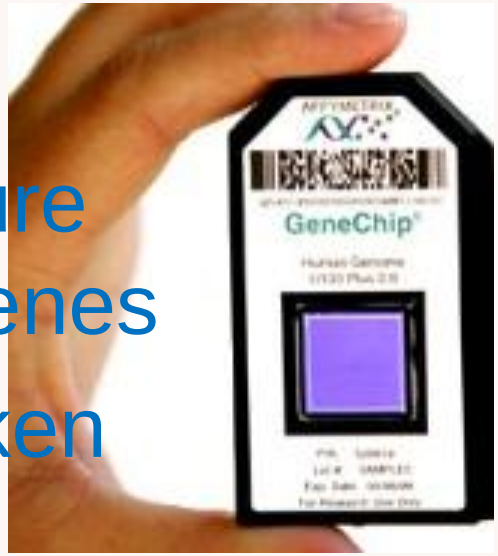
Mycoplasma

- Tiny bacteria which routinely infect microbiology laboratories
- Not easy to detect
- Mycoplasma infection makes sample measurements useless
- Mycoplasma infects 10-25% laboratory cultures. (Variable but high).

mycoplasma capricolum



Affymetrix HG-U133 +2



- First single microarray to measure RNA expression of all human genes
- Design based on sequences taken from Human reference genome GenBank, dbEST, RefSeq (UniGene build 133, April 2001)
- HG-U133 +2 also includes expressed sequence tags (ESTs)
- Typically 11 measurements (probes) per DNA sequence

HG-U133 +2 probeset 1570561_at

- Affymetrix microarray HG-U133 +2 probeset 1570561_at was derived from GenBank AF241217
- AF241217 “Homo sapiens unknown sequence” was submitted to GenBank in 2000

Evidence: Blast

- Blast used to compare AF241217 DNA sequence with all sequenced species
- AF241217 sequence matches itself and various species of Mycoplasma

EBI > Tools > Sequence Similarity Searching > NCBI BLAST

NCBI BLAST Results

Summary Table | Tool Output | Visual Output | Submission Details | Submit Another Job

Alignments

Selection: Show Annotations | Hide Annotations | Show Alignments | Hide Alignments

Download in fasta format

Clear Selection | Select All | Invert Selection

Align.	DB:ID	Source	Length	Score	Identities	E()
☑ 1	EM_HTG:AF241217	Homo sapiens unknown sequence. *Cross-references and related information in:* ▶Ontologies	249	225	100.0	1.0E-121
☑ 2	EM_PRO:FJ876260	Mycoplasma orale strain MT-4 16S ribosomal RNA gene, partial sequence; 16S-23S ribosomal RNA intergenic spacer, complete sequence; and 23S ribosomal RNA gene, partial sequence. *Cross-references and related information in:* ▶Ontologies	2231	218	100.0	1.0E-117
☑ 3	EM_PRO:AF294995	Mycoplasma orale 16S-23S ribosomal RNA intergenic spacer, complete sequence; and 23S ribosomal RNA gene, partial sequence. *Cross-references and related information in:* ▶Literature ▶Ontologies	738	218	100.0	1.0E-117
☑ 4	EM_PRO:JN689375	Mycoplasma orale isolate LJH 16S ribosomal RNA gene, partial sequence; 16S-23S ribosomal RNA intergenic spacer, complete sequence; and 23S ribosomal RNA gene, partial sequence. *Cross-references and related information in:* ▶Ontologies	535	214	99.0	1.0E-115
☑ 5	EM_PRO:AY737010	Mycoplasma orale 16S ribosomal RNA gene, partial sequence; 16S-23S ribosomal RNA intergenic spacer, complete sequence; and 23S ribosomal RNA gene, partial sequence. *Cross-references and related information in:* ▶Literature ▶Ontologies	882	214	99.0	1.0E-115
☑ 6	EM_PRO:AY762640	Mycoplasma indiense 16S ribosomal RNA gene, partial sequence; 16S-23S ribosomal RNA intergenic spacer, complete sequence; and 23S ribosomal RNA gene, partial sequence.	870	207	99.0	1.0E-111

HG-U133 +2 probeset 1570561_at from Mycoplasma?

- Matches 16S-23S rRNA intergenic spacer (ITS) which is already used to detect Mycoplasma.
- No similarities with any human transcript or genome sequence
- AF241217 came from Mycoplasma contaminated human cell line

1570561_at from Mycoplasma?

- None of the other ~47,400 complete sequence targeted by HG-U133 +2 matches Mycoplasma arthritidis

Evidence: Published gene expression data

- In thousands of data from published peer-reviewed journal articles, the 1570561_at gene is expressed where contamination by Mycoplasma might be expected.
- Yes. 1570561_at is expressed in cultured cells. (Ie cells from microbiology laboratories rather than biopsies or tissue samples from patients).

Gene Expression Omnibus

- NCBI GEO is an archive containing tens of thousands of gene expression datasets.
- All HG-133 +2 datasets were loaded into RNAnet in February 2007 (total 2757 samples)
- RNAnet allows instant access to normalised microarray data

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).

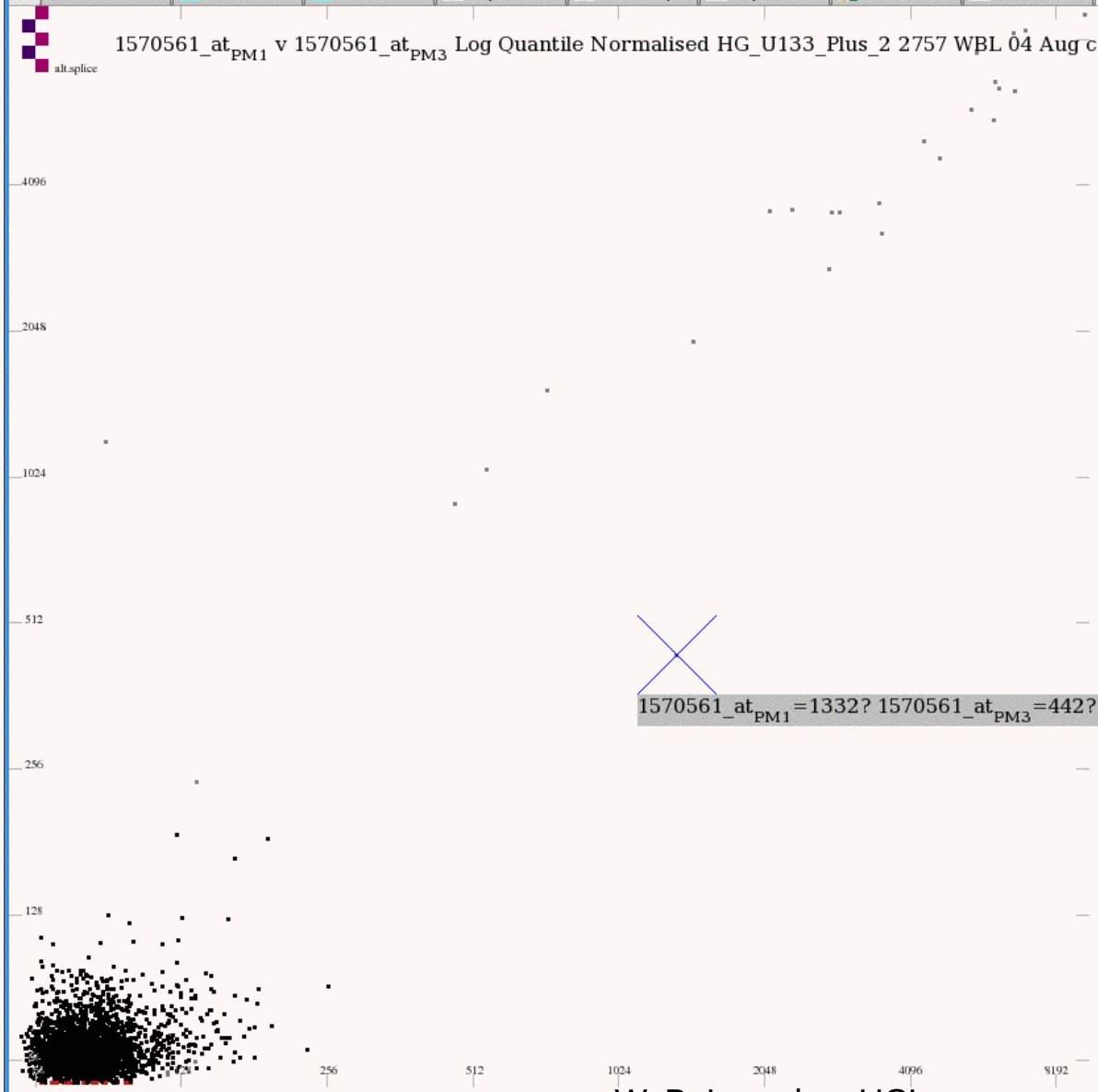
Expression of 1570561_at in GEO

Scatter plot of Human Genes - Mozilla Firefox

File Edit View History Bookmarks Tools Help

bioinformatics.essex.ac.uk/users/wlangdon/rnanet/scatter.html#1570561_at.pm1,1570561_at.pm3

Gmail - I... | William L... | FASTA R... | http...1217 | ENA Seq... | Expressi... | 16s-23s ... | BioTech... | Scatt... | Gene Ex... | JMB :: Jo... | App



Expression of Human Genes

1570561_at ☐ MM/PM 1

1570561_at ☐ MM/PM 3 probes

plot clear

resize (also clears)

Example

top

StatCounter.com

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

Gmail - I... | William L... | FASTA R... | http...1217 | ENA Seq... | Expressi... | 16s-23s ... | BioTech... | GEO ... | Gene Ex... | JMB :: Jo... | App

NCBI

Gene Expression Omnibus

HOME SEARCH SITE MAP | GEO Publications | FAQ | MIAME | Email

NCBI > GEO > **Accession Display**

Not logged in | L

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

Sample GSM48672 Query DataSets for 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

Contamination in other direction Human genes → other species

- Many human genes in non-primate DNA sequence databases

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
  - 16,923 organisms have already been sequenced (RefSeq March 2012).
- Known problem. Nobody working on a solution? Will only get worse.
- So what?
- "Due dilligence". Can't take most important bioinformatics database on trust

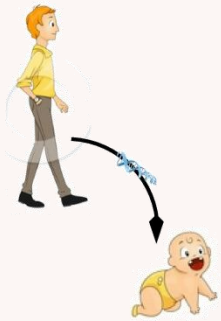
Genes Spread

- 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.

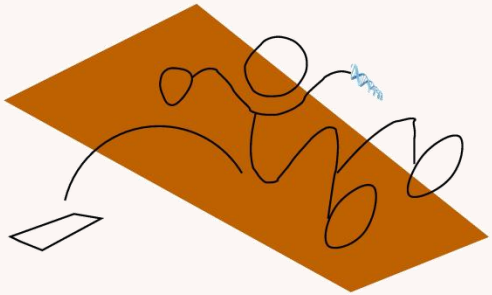
Summary

- HG-U133 +2 probeset 1570561_at originates from mycoplasma not humans.
- 1570561_at may detect mycoplasma RNA in human microarray sample.
- ≈1% of GEO database compromised.
- Abundant human DNA contamination identified in non-primate genome databases.
- Found 2 non-human cases → others
- Problems reported but not fixed.

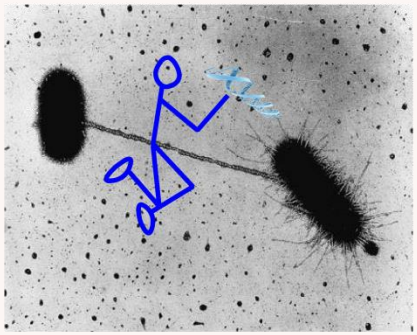
Genes Jump Silicon Barrier



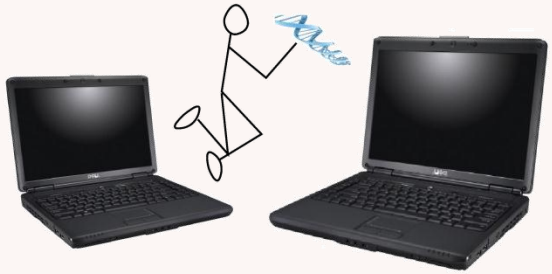
Mendel 1865



Jumping genes
McClintock 1930



Horizontal gene transfer
1959



Gene transfer to GenBank
Today

- 1865 vertical gene transfer
- 1930 gene transfer along chromosomes
- 1959 antibiotic resistance between species
- Jumping genes escape biology, cross the silicon barrier and roam computer databases

END

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

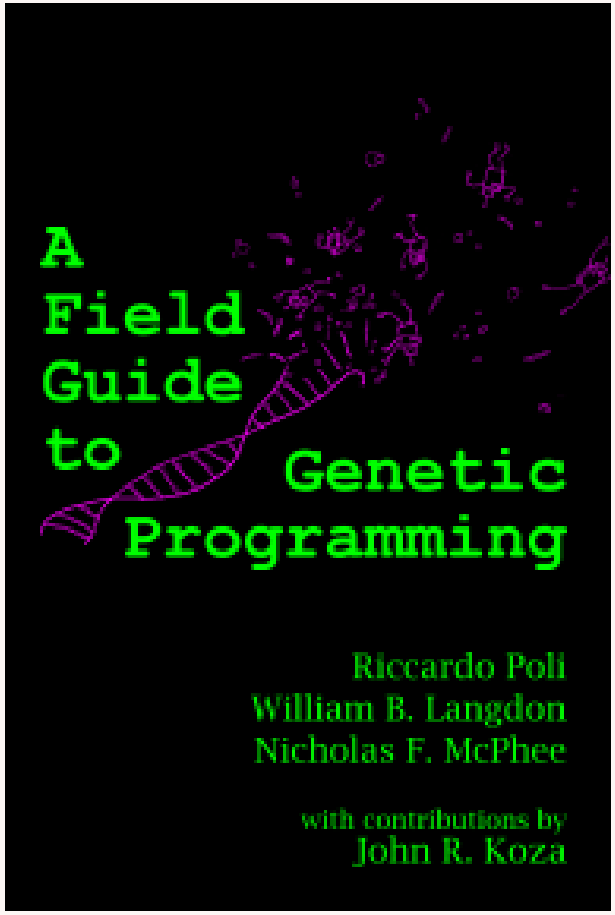
Mycoplasma genes in the Human Genome

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)

A Field Guide To Genetic Programming

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



Free PDF

The Genetic Programming Bibliography

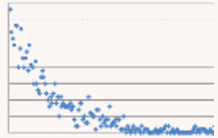
The largest, most complete, collection of GP papers.
http://www.cs.bham.ac.uk/~wbl/biblio/

With 7,878 references, and 6,250 online publications, the GP Bibliography is a vital resource to the computer science, artificial intelligence, machine learning, and evolutionary computing communities.

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

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

A personalised list of every author's GP publications.



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