Challenges in Running a Small Bioinformatics Service Group

Carl Virtanen, Manager of Bioinformatics, OCI

Overview of Group at OCI

- Housed at the OCI Genomics Centre (formerly UHN MAC)
- 1 full time manager/bioinformatician, 1 full time DBA and support bioinformatician, 1-3 co-op students, 1 data analyst
- Small cluster (~100 cores, 40 TB space) and high end desktops
- Budget for group is “floating” (grants when available, chargeback for services, institutional support)
- Has been in existence for 10 years
- Currently undergoing a big expansion of both hardware and personnel to meet next-gen sequencing needs

Overview continued...

- Average of about 70 full data analyses per year
- Collaboration and fee-for-service analyses
- Institutional demands (infrastructure, committees, planning, grant writing, publications, etc)
- Websites, DB's, custom programming, sys admin etc
- Sequencing is ramping up
- For a small group this adds up...

Problem Areas

- Staffing is always a problem (everyone has to be able to do bits of everything)
- Pricing (and associated billing, invoicing etc overhead)
- Fluidity of databases and versioning of arrays and genomes is a major problem (eg naming conventions)
- Approaches to, and what is considered “best practice”, analysis (TMTOWTDI)
- Communicating and meeting expectations for customers and collaborators is the hardest area to deal with

Our Solutions

- Project based pricing works the best (as opposed to billable hrs)
- 2 week turnaround (max).
- Standardization of workflow for technologies (expression, CNV, SNP, ChIP/Chip, sequencing)
- Standardized reports. Could be dropped into a publication (authorship at customers discretion ie-usually none!)
- Keeping up to date with literature and analysis "trends"
- Stick to "hard stats" and results. Stay away from more interpretive/interactive types of results (e.g. GO=good, pathway analysis=trouble) except for collaborative work. Offer software advice and "prep" for downstream work (Cytoscape, GSEA, DAVID etc)

Managing Throughput

- Communication from the very beginning (pre-experimental design) is key to success
- Knowing the biology and speaking the language (reduces anxiety, helps with the interpretation of large results sets).
- Interpretation is still up to the customer. Just give the facts.
- TIMELINES!
- Build reusable tools that will aid internal workflows, showcase (advertising) and not reinvent wheels (unless its a better wheel)

Some Problems in Dealing with "Gene" Level Data on Arrays

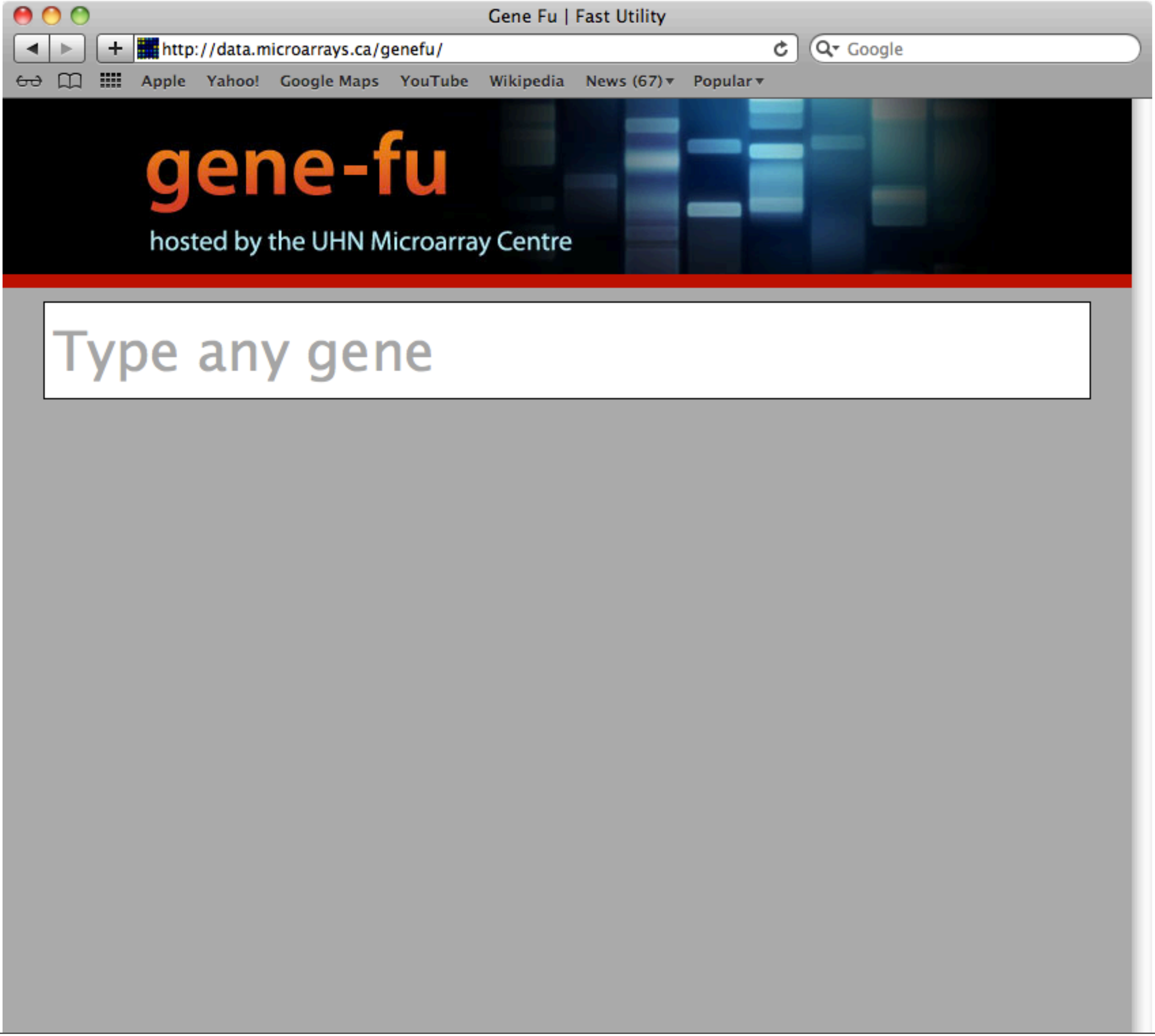
- Naming conventions are at odds with common usage
- Looking up information on genes when going through big lists is a bit slow (multiple clicks) when you're sitting down with somebody interactively showing them results
- Positions of elements on the genome change (or elements even disappear!) over time
- For microarrays, the above two items end up causing major headaches

Dealing with Naming Conventions

- Build an easy to use tool!

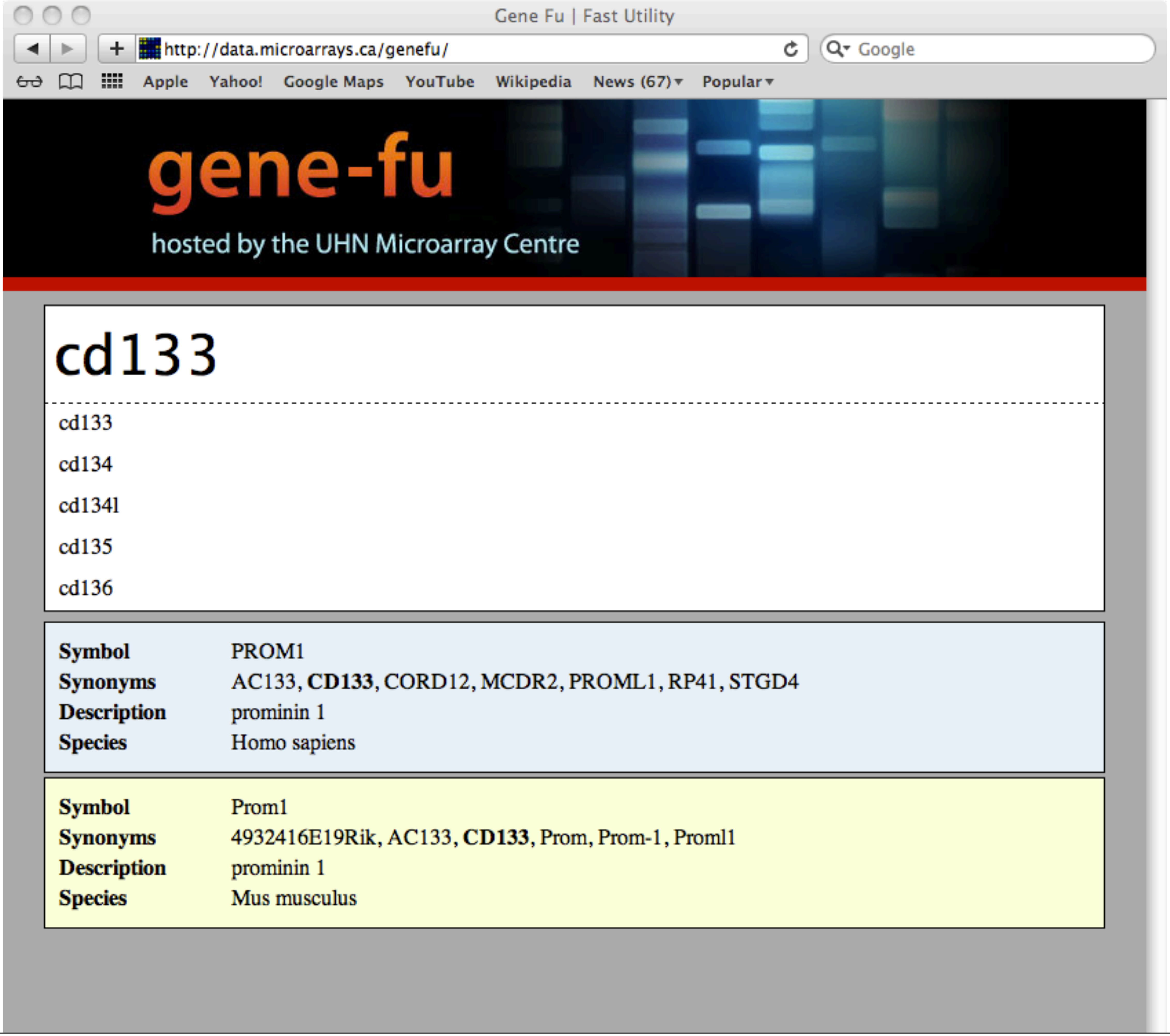
Genefu

- Searches symbols, synonyms, and full names
- Mouse and human only
- Auto-complete in case you forget
- Just the main points (with links if you really need to know more)
- Simple design
- Reusable



gene-fu

hosted by the UHN Microarray Centre



gene-fu

hosted by the UHN Microarray Centre

cd133

cd133

cd134

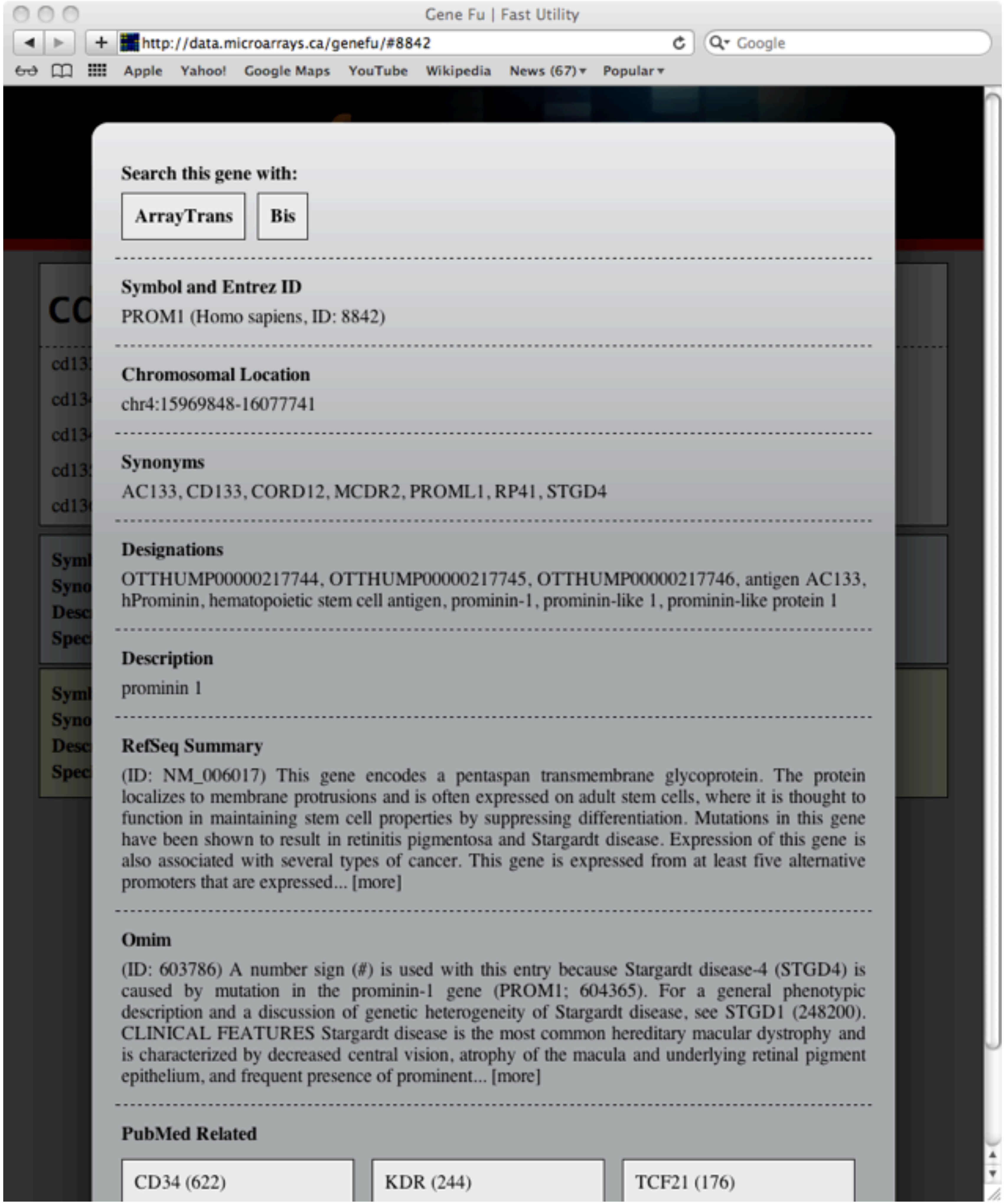
cd1341

cd135

cd136

Symbol	PROM1
Synonyms	AC133, **CD133**, CORD12, MCDR2, PROML1, RP41, STGD4
Description	prominin 1
Species	Homo sapiens

Symbol	Prom1
Synonyms	4932416E19Rik, AC133, **CD133**, Prom, Prom-1, Proml1
Description	prominin 1
Species	Mus musculus



Search this gene with:

ArrayTrans | Bis

Symbol and Entrez ID

PROM1 (Homo sapiens, ID: 8842)

Chromosomal Location

chr4:15969848-16077741

Synonyms

AC133, CD133, CORD12, MCDR2, PROML1, RP41, STGD4

Designations

OTTHUMP00000217744, OTTHUMP00000217745, OTTHUMP00000217746, antigen AC133, hProminin, hematopoietic stem cell antigen, prominin-1, prominin-like 1, prominin-like protein 1

Description

prominin 1

RefSeq Summary

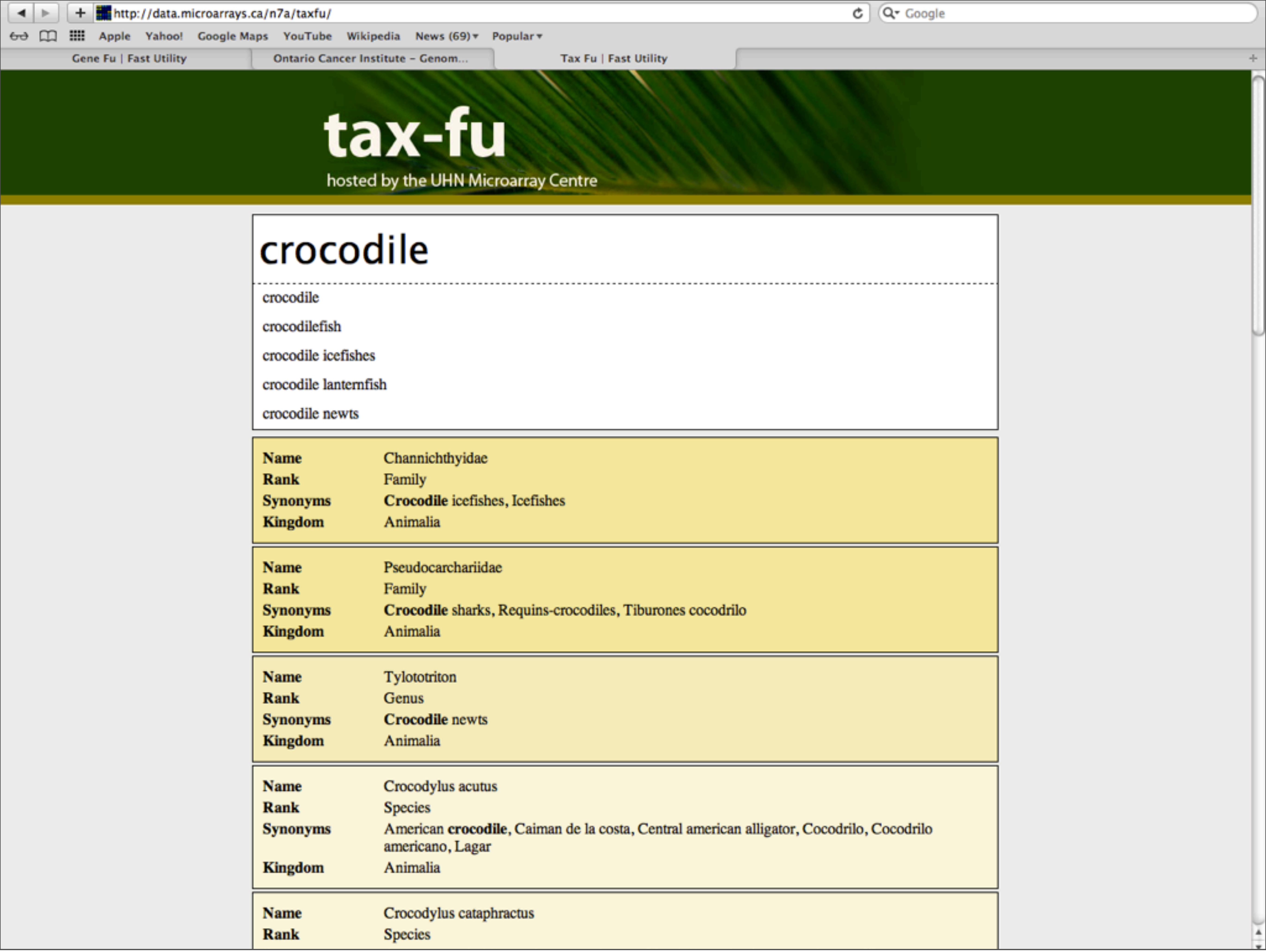
(ID: NM_006017) This gene encodes a pentaspan transmembrane glycoprotein. The protein localizes to membrane protrusions and is often expressed on adult stem cells, where it is thought to function in maintaining stem cell properties by suppressing differentiation. Mutations in this gene have been shown to result in retinitis pigmentosa and Stargardt disease. Expression of this gene is also associated with several types of cancer. This gene is expressed from at least five alternative promoters that are expressed... [more]

Omim

(ID: 603786) A number sign (#) is used with this entry because Stargardt disease-4 (STGD4) is caused by mutation in the prominin-1 gene (PROM1; 604365). For a general phenotypic description and a discussion of genetic heterogeneity of Stargardt disease, see STGD1 (248200). CLINICAL FEATURES Stargardt disease is the most common hereditary macular dystrophy and is characterized by decreased central vision, atrophy of the macula and underlying retinal pigment epithelium, and frequent presence of prominent... [more]

PubMed Related

CD34 (622) | KDR (244) | TCF21 (176)



tax-fu

hosted by the UHN Microarray Centre

crocodile

crocodile

crocodilefish

crocodile icefishes

crocodile lanternfish

crocodile newts

Name	Channichthyidae
Rank	Family
Synonyms	**Crocodile** icefishes, Icefishes
Kingdom	Animalia

Name	Pseudocarchariidae
Rank	Family
Synonyms	**Crocodile** sharks, Requins-crocodiles, Tiburones cocodrilo
Kingdom	Animalia

Name	Tylototriton
Rank	Genus
Synonyms	**Crocodile** newts
Kingdom	Animalia

Name	Crocodylus acutus
Rank	Species
Synonyms	American **crocodile**, Caiman de la costa, Central american alligator, Cocodrilo, Cocodrilo americano, Lagar
Kingdom	Animalia

Name	Crocodylus cataphractus
Rank	Species

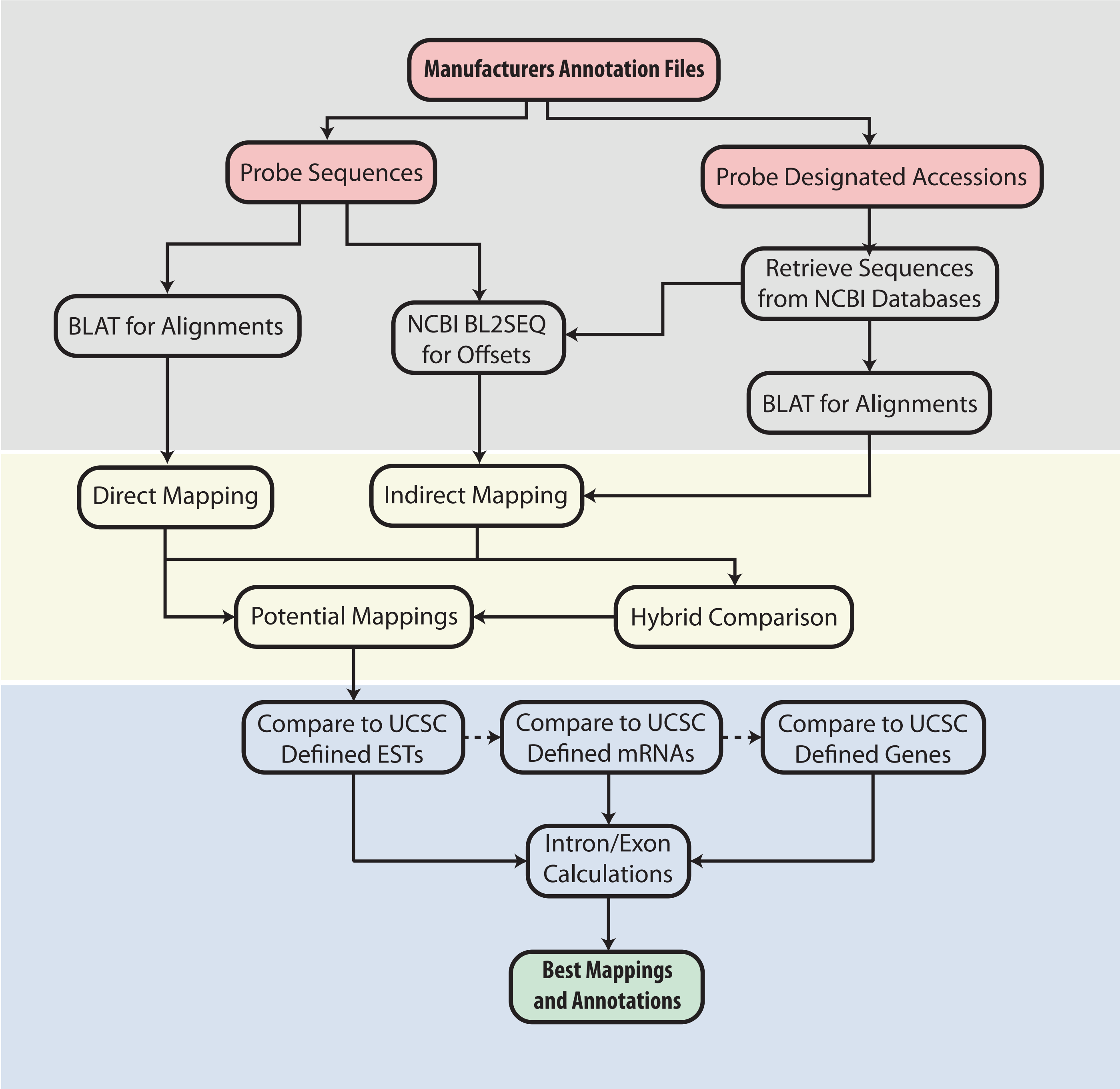
Array Specific Problems

- Customers almost always want to compare across arrays or array versions to (un)published data (or have "a list")
- Customers almost always use a common usage naming convention
- EVERY array manufacturer has problematic probes (probes on introns, probes on chimeric sequences, probes to the wrong species even!)
- Build an easy to use tool!

Arraytrans

- Assume that one thing is absolutely correct: the sequence of the probe
- Name may be wrong or changed
- Sequence it was designed to may have disappeared/deprecated
- Once everything is scaffolded to the same build of a genome and its associated databases, searching and cross-matching is (somewhat) easy using sql joins

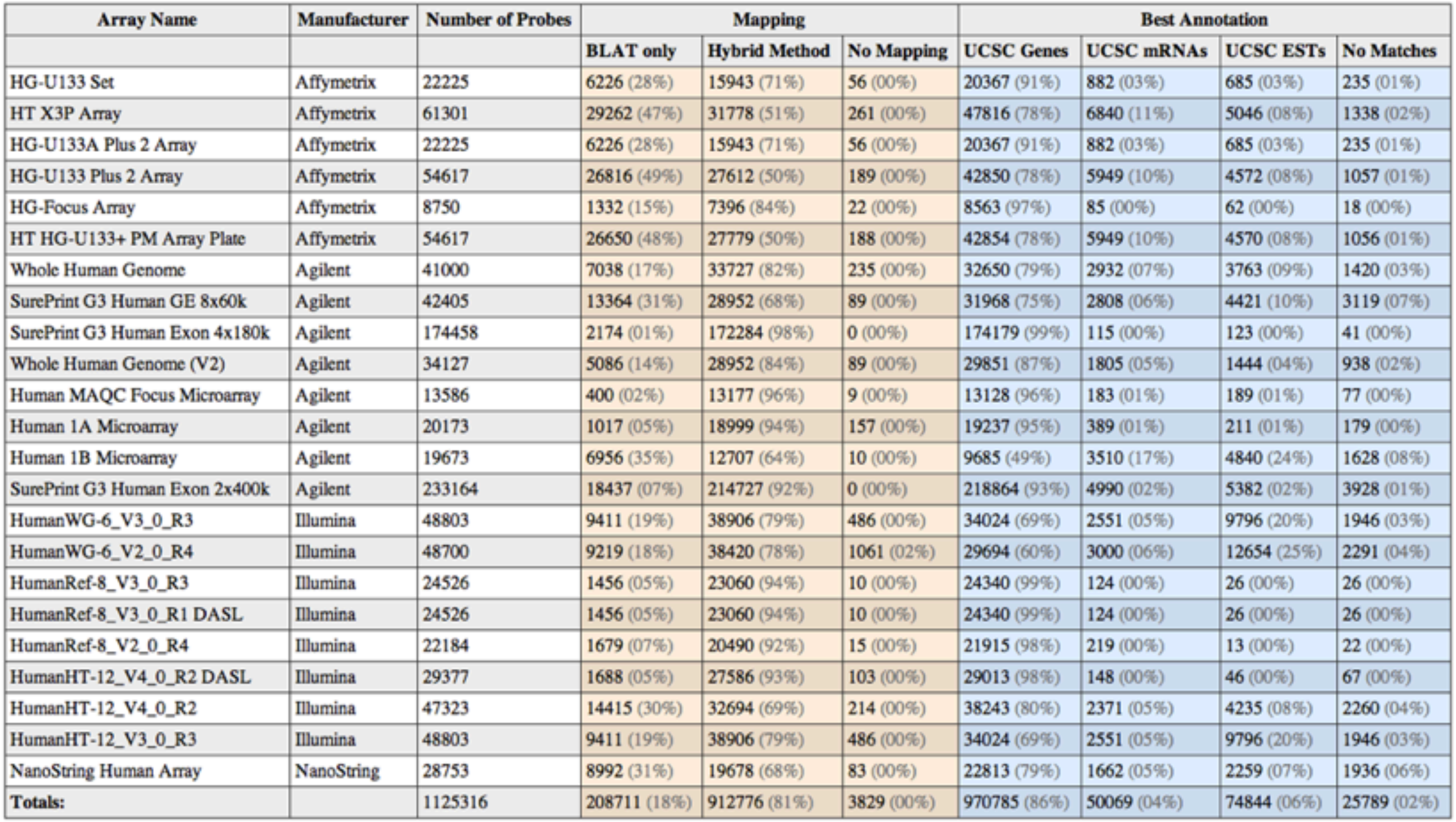
Arraytrans Overview



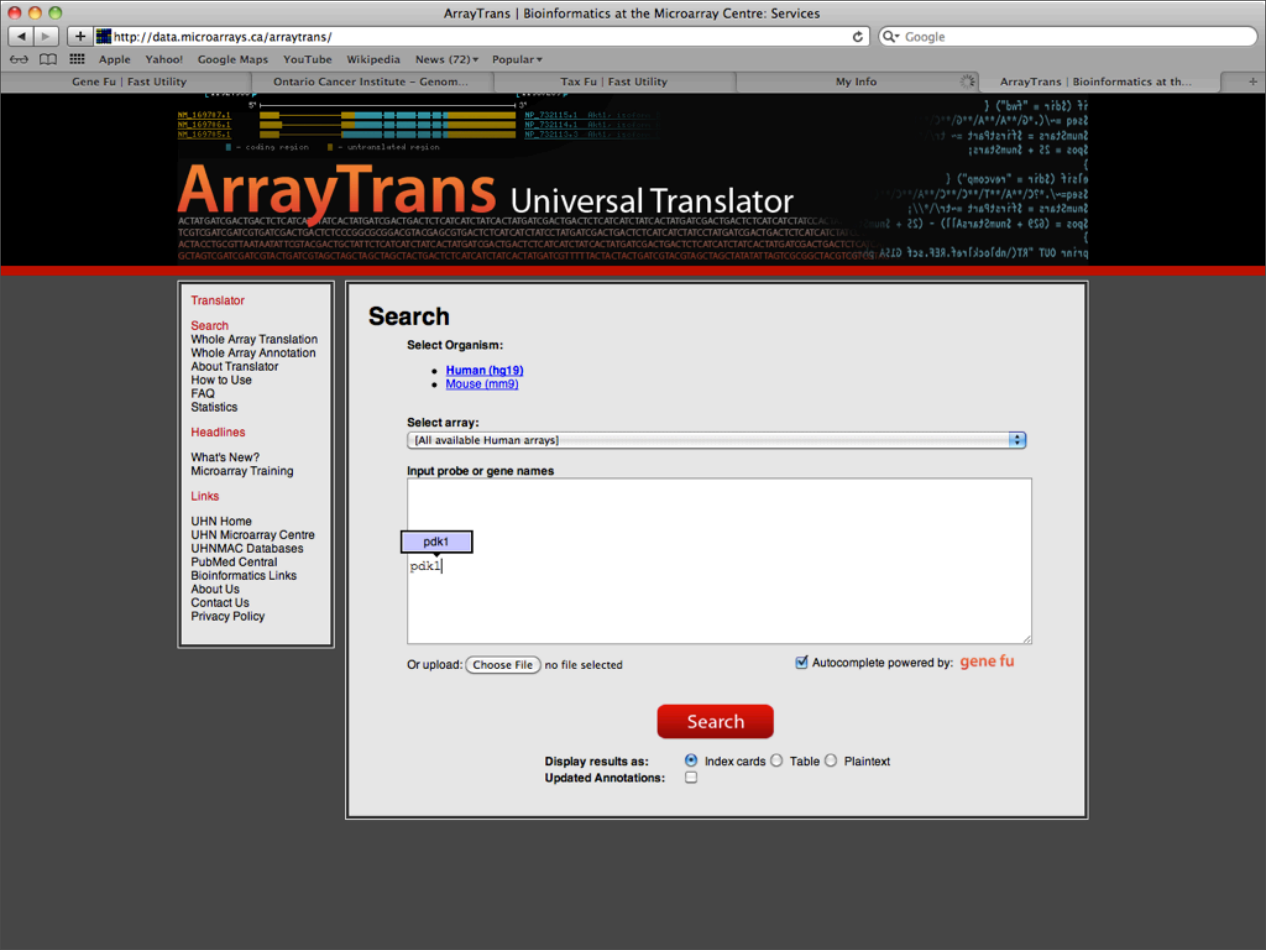
Manufacturers Annotation Files
Probe Sequences
Probe Designated Accessions
Retrieve Sequences from NCBI Databases
BLAT for Alignments
NCBI BL2SEQ for Offsets
BLAT for Alignments
Direct Mapping
Indirect Mapping
Potential Mappings
Hybrid Comparison
Compare to UCSC Defined ESTs
Compare to UCSC Defined mRNAs
Compare to UCSC Defined Genes
Intron/Exon Calculations
Best Mappings and Annotations

Tuesday, November 27, 2012

Re-annotation Statistics



Array Name	Manufacturer	Number of Probes	Mapping			Best Annotation			
			BLAT only	Hybrid Method	No Mapping	UCSC Genes	UCSC mRNAs	UCSC ESTs	No Matches
HG-U133 Set	Affymetrix	22225	6226 (28%)	15943 (71%)	56 (00%)	20367 (91%)	882 (03%)	685 (03%)	235 (01%)
HT X3P Array	Affymetrix	61301	29262 (47%)	31778 (51%)	261 (00%)	47816 (78%)	6840 (11%)	5046 (08%)	1338 (02%)
HG-U133A Plus 2 Array	Affymetrix	22225	6226 (28%)	15943 (71%)	56 (00%)	20367 (91%)	882 (03%)	685 (03%)	235 (01%)
HG-U133 Plus 2 Array	Affymetrix	54617	26816 (49%)	27612 (50%)	189 (00%)	42850 (78%)	5949 (10%)	4572 (08%)	1057 (01%)
HG-Focus Array	Affymetrix	8750	1332 (15%)	7396 (84%)	22 (00%)	8563 (97%)	85 (00%)	62 (00%)	18 (00%)
HT HG-U133+ PM Array Plate	Affymetrix	54617	26650 (48%)	27779 (50%)	188 (00%)	42854 (78%)	5949 (10%)	4570 (08%)	1056 (01%)
Whole Human Genome	Agilent	41000	7038 (17%)	33727 (82%)	235 (00%)	32650 (79%)	2932 (07%)	3763 (09%)	1420 (03%)
SurePrint G3 Human GE 8x60k	Agilent	42405	13364 (31%)	28952 (68%)	89 (00%)	31968 (75%)	2808 (06%)	4421 (10%)	3119 (07%)
SurePrint G3 Human Exon 4x180k	Agilent	174458	2174 (01%)	172284 (98%)	0 (00%)	174179 (99%)	115 (00%)	123 (00%)	41 (00%)
Whole Human Genome (V2)	Agilent	34127	5086 (14%)	28952 (84%)	89 (00%)	29851 (87%)	1805 (05%)	1444 (04%)	938 (02%)
Human MAQC Focus Microarray	Agilent	13586	400 (02%)	13177 (96%)	9 (00%)	13128 (96%)	183 (01%)	189 (01%)	77 (00%)
Human 1A Microarray	Agilent	20173	1017 (05%)	18999 (94%)	157 (00%)	19237 (95%)	389 (01%)	211 (01%)	179 (00%)
Human 1B Microarray	Agilent	19673	6956 (35%)	12707 (64%)	10 (00%)	9685 (49%)	3510 (17%)	4840 (24%)	1628 (08%)
SurePrint G3 Human Exon 2x400k	Agilent	233164	18437 (07%)	214727 (92%)	0 (00%)	218864 (93%)	4990 (02%)	5382 (02%)	3928 (01%)
HumanWG-6_V3_0_R3	Illumina	48803	9411 (19%)	38906 (79%)	486 (00%)	34024 (69%)	2551 (05%)	9796 (20%)	1946 (03%)
HumanWG-6_V2_0_R4	Illumina	48700	9219 (18%)	38420 (78%)	1061 (02%)	29694 (60%)	3000 (06%)	12654 (25%)	2291 (04%)
HumanRef-8_V3_0_R3	Illumina	24526	1456 (05%)	23060 (94%)	10 (00%)	24340 (99%)	124 (00%)	26 (00%)	26 (00%)
HumanRef-8_V3_0_R1 DASL	Illumina	24526	1456 (05%)	23060 (94%)	10 (00%)	24340 (99%)	124 (00%)	26 (00%)	26 (00%)
HumanRef-8_V2_0_R4	Illumina	22184	1679 (07%)	20490 (92%)	15 (00%)	21915 (98%)	219 (00%)	13 (00%)	22 (00%)
HumanHT-12_V4_0_R2 DASL	Illumina	29377	1688 (05%)	27586 (93%)	103 (00%)	29013 (98%)	148 (00%)	46 (00%)	67 (00%)
HumanHT-12_V4_0_R2	Illumina	47323	14415 (30%)	32694 (69%)	214 (00%)	38243 (80%)	2371 (05%)	4235 (08%)	2260 (04%)
HumanHT-12_V3_0_R3	Illumina	48803	9411 (19%)	38906 (79%)	486 (00%)	34024 (69%)	2551 (05%)	9796 (20%)	1946 (03%)
NanoString Human Array	NanoString	28753	8992 (31%)	19678 (68%)	83 (00%)	22813 (79%)	1662 (05%)	2259 (07%)	1936 (06%)
Totals:		1125316	208711 (18%)	912776 (81%)	3829 (00%)	970785 (86%)	50069 (04%)	74844 (06%)	25789 (02%)



Search

Select Organism:

- Human (hg19)
- Mouse (mm9)

Select array: [All available Human arrays]

Input probe or gene names

pdk1

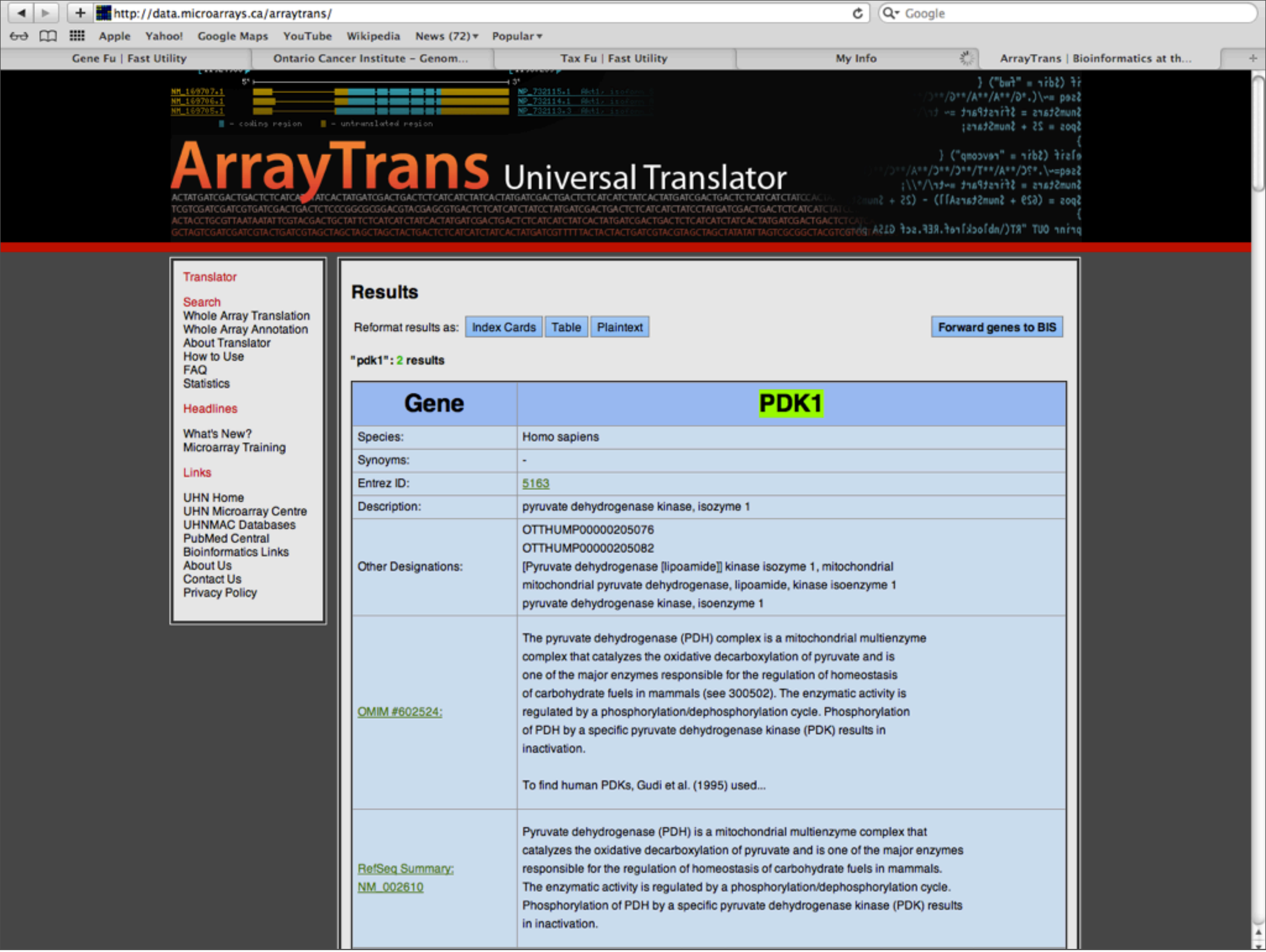
pdk1

Or upload: Choose File no file selected

☑ Autocomplete powered by: gene fu

Display results as: ● Index cards ○ Table ○ Plaintext

Updated Annotations: □


http://data.microarrays.ca/arraytrans/
Google
Apple
Yahoo!
Google Maps
YouTube
Wikipedia
News (72)
Popular
Gene Fu | Fast Utility
Ontario Cancer Institute – Genom...
Tax Fu | Fast Utility
My Info
ArrayTrans | Bioinformatics at th...
ArrayTrans Universal Translator

Translator

Search
Whole Array Translation
Whole Array Annotation
About Translator
How to Use
FAQ
Statistics

Headlines

What's New?
Microarray Training

Links

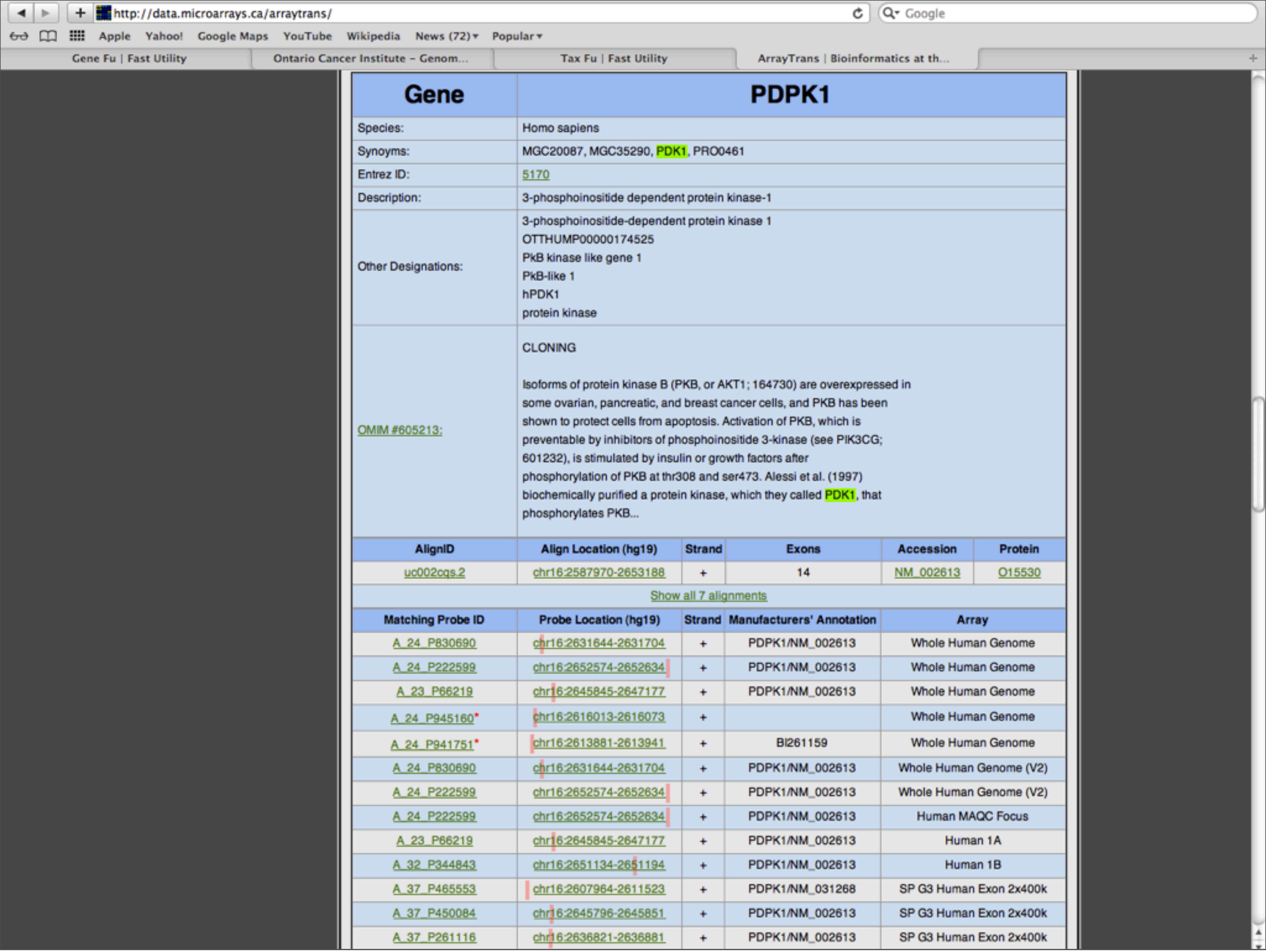
UHN Home
UHN Microarray Centre
UHNMAC Databases
PubMed Central
Bioinformatics Links
About Us
Contact Us
Privacy Policy

Results

Reformat results as: Index Cards Table Plaintext Forward genes to BIS

"pdk1": 2 results

Gene	PDK1
Species:	Homo sapiens
Synoyms:	-
Entrez ID:	5163
Description:	pyruvate dehydrogenase kinase, isozyme 1
Other Designations:	OTTHUMP00000205076 OTTHUMP00000205082 [Pyruvate dehydrogenase [lipoamide]] kinase isozyme 1, mitochondrial mitochondrial pyruvate dehydrogenase, lipoamide, kinase isoenzyme 1 pyruvate dehydrogenase kinase, isoenzyme 1
OMIM #602524:	The pyruvate dehydrogenase (PDH) complex is a mitochondrial multienzyme complex that catalyzes the oxidative decarboxylation of pyruvate and is one of the major enzymes responsible for the regulation of homeostasis of carbohydrate fuels in mammals (see 300502). The enzymatic activity is regulated by a phosphorylation/dephosphorylation cycle. Phosphorylation of PDH by a specific pyruvate dehydrogenase kinase (PDK) results in inactivation. To find human PDKs, Gudi et al. (1995) used...
RefSeq Summary: NM_002610	Pyruvate dehydrogenase (PDH) is a mitochondrial multienzyme complex that catalyzes the oxidative decarboxylation of pyruvate and is one of the major enzymes responsible for the regulation of homeostasis of carbohydrate fuels in mammals. The enzymatic activity is regulated by a phosphorylation/dephosphorylation cycle. Phosphorylation of PDH by a specific pyruvate dehydrogenase kinase (PDK) results in inactivation.

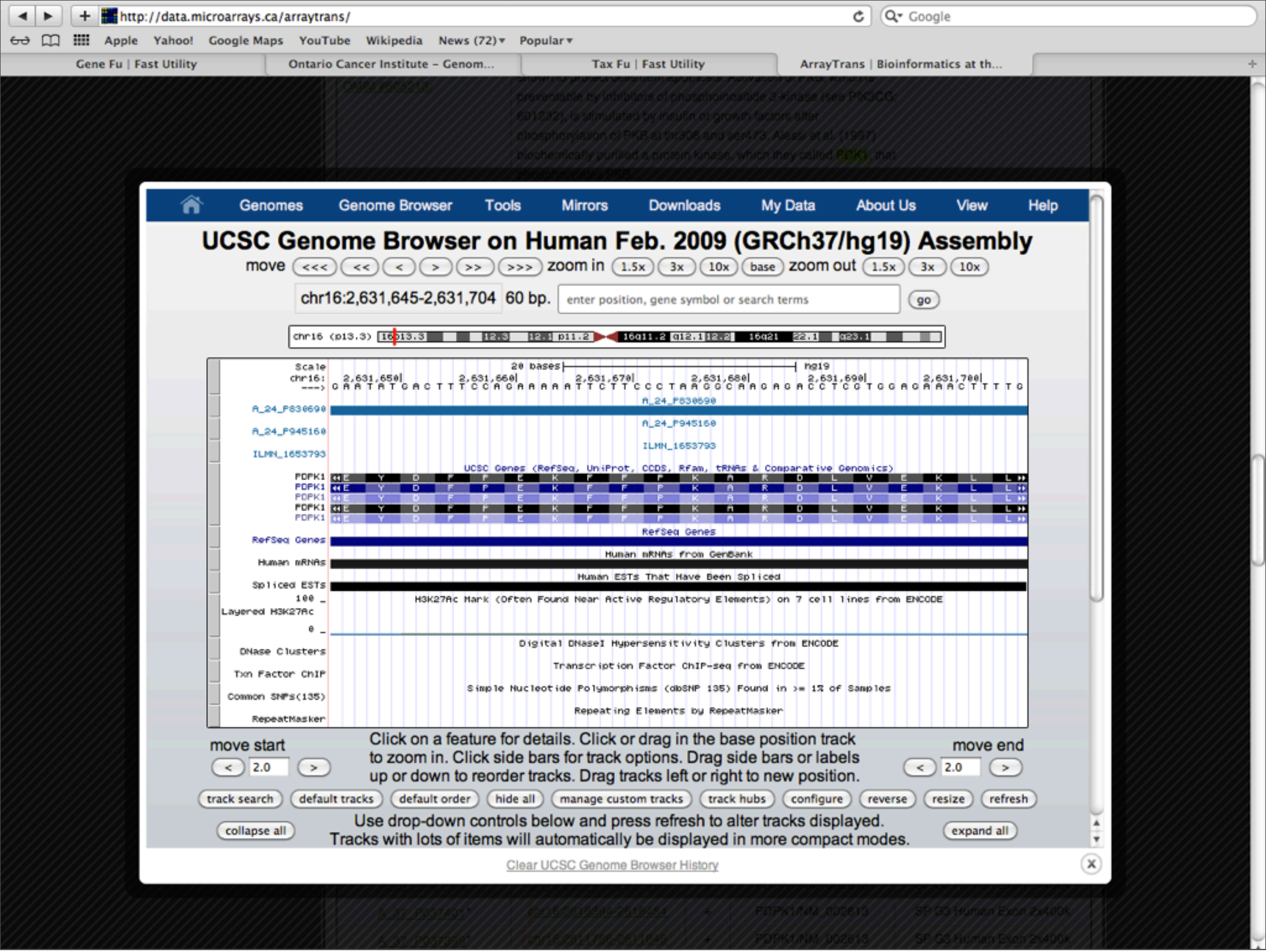

http://data.microarrays.ca/arraytrans/
Google
Apple
Yahoo!
Google Maps
YouTube
Wikipedia
News (72)
Popular
Gene Fu | Fast Utility
Ontario Cancer Institute – Genom...
Tax Fu | Fast Utility
ArrayTrans | Bioinformatics at th...

Gene	PDPK1
Species:	Homo sapiens
Synoyms:	MGC20087, MGC35290, PDK1, PRO0461
Entrez ID:	5170
Description:	3-phosphoinositide dependent protein kinase-1
Other Designations:	3-phosphoinositide-dependent protein kinase 1 OTTHUMP00000174525 PkB kinase like gene 1 PkB-like 1 hPDK1 protein kinase
OMIM #605213:	CLONING Isoforms of protein kinase B (PKB, or AKT1; 164730) are overexpressed in some ovarian, pancreatic, and breast cancer cells, and PKB has been shown to protect cells from apoptosis. Activation of PKB, which is preventable by inhibitors of phosphoinositide 3-kinase (see PIK3CG; 601232), is stimulated by insulin or growth factors after phosphorylation of PKB at thr308 and ser473. Alessi et al. (1997) biochemically purified a protein kinase, which they called PDK1, that phosphorylates PKB...

AlignID	Align Location (hg19)	Strand	Exons	Accession	Protein
uc002cqs.2	chr16:2587970-2653188	+	14	NM_002613	O15530

Show all 7 alignments

Matching Probe ID	Probe Location (hg19)	Strand	Manufacturers' Annotation	Array
A_24_P830690	chr16:2631644-2631704	+	PDPK1/NM_002613	Whole Human Genome
A_24_P222599	chr16:2652574-2652634	+	PDPK1/NM_002613	Whole Human Genome
A_23_P66219	chr16:2645845-2647177	+	PDPK1/NM_002613	Whole Human Genome
A_24_P945160*	chr16:2616013-2616073	+		Whole Human Genome
A_24_P941751*	chr16:2613881-2613941	+	BI261159	Whole Human Genome
A_24_P830690	chr16:2631644-2631704	+	PDPK1/NM_002613	Whole Human Genome (V2)
A_24_P222599	chr16:2652574-2652634	+	PDPK1/NM_002613	Whole Human Genome (V2)
A_24_P222599	chr16:2652574-2652634	+	PDPK1/NM_002613	Human MAQC Focus
A_23_P66219	chr16:2645845-2647177	+	PDPK1/NM_002613	Human 1A
A_32_P344843	chr16:2651134-2651194	+	PDPK1/NM_002613	Human 1B
A_37_P465553	chr16:2607964-2611523	+	PDPK1/NM_031268	SP G3 Human Exon 2x400k
A_37_P450084	chr16:2645796-2645851	+	PDPK1/NM_002613	SP G3 Human Exon 2x400k
A_37_P261116	chr16:2636821-2636881	+	PDPK1/NM_002613	SP G3 Human Exon 2x400k



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr16:2,631,645-2,631,704 60 bp.

enter position, gene symbol or search terms
go
chr16 (p13.3)
16p13.3
12.3
12.1
p11.2
16q11.2
q12.1
12.2
16q21
22.1
q23.1
Scale
chr16:
20 bases
hg19
2,631,650
2,631,660
2,631,670
2,631,680
2,631,690
2,631,700
--->
G A A T A T G A C T T T C C A G A A A A A T T C T T C C C T A A G G C A A G A G A C C T C G T G G A G A A A C T T T T G
A_24_P830690
A_24_P945160
ILMN_1653793
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
PDPK1
PDPK1
PDPK1
PDPK1
PDPK1
E Y D F P E K F F P K A R D L V E K L L
RefSeq Genes
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
100 _
Layered H3K27Ac
0 _
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
Digital DNaseI Hypersensitivity Clusters from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
Common SNPs(135)
Simple Nucleotide Polymorphisms (dbSNP 135) Found in >= 1% of Samples
RepeatMasker
Repeating Elements by RepeatMasker
move start
<
2.0
>
move end
<
2.0
>

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

Summary

- Proper communication is paramount to operating successfully
- Time resource management is critical
- Keep focused but allow time to develop new and helpful tools which prevents customer "burnout" and keeps you up-to-date on technologies and issues

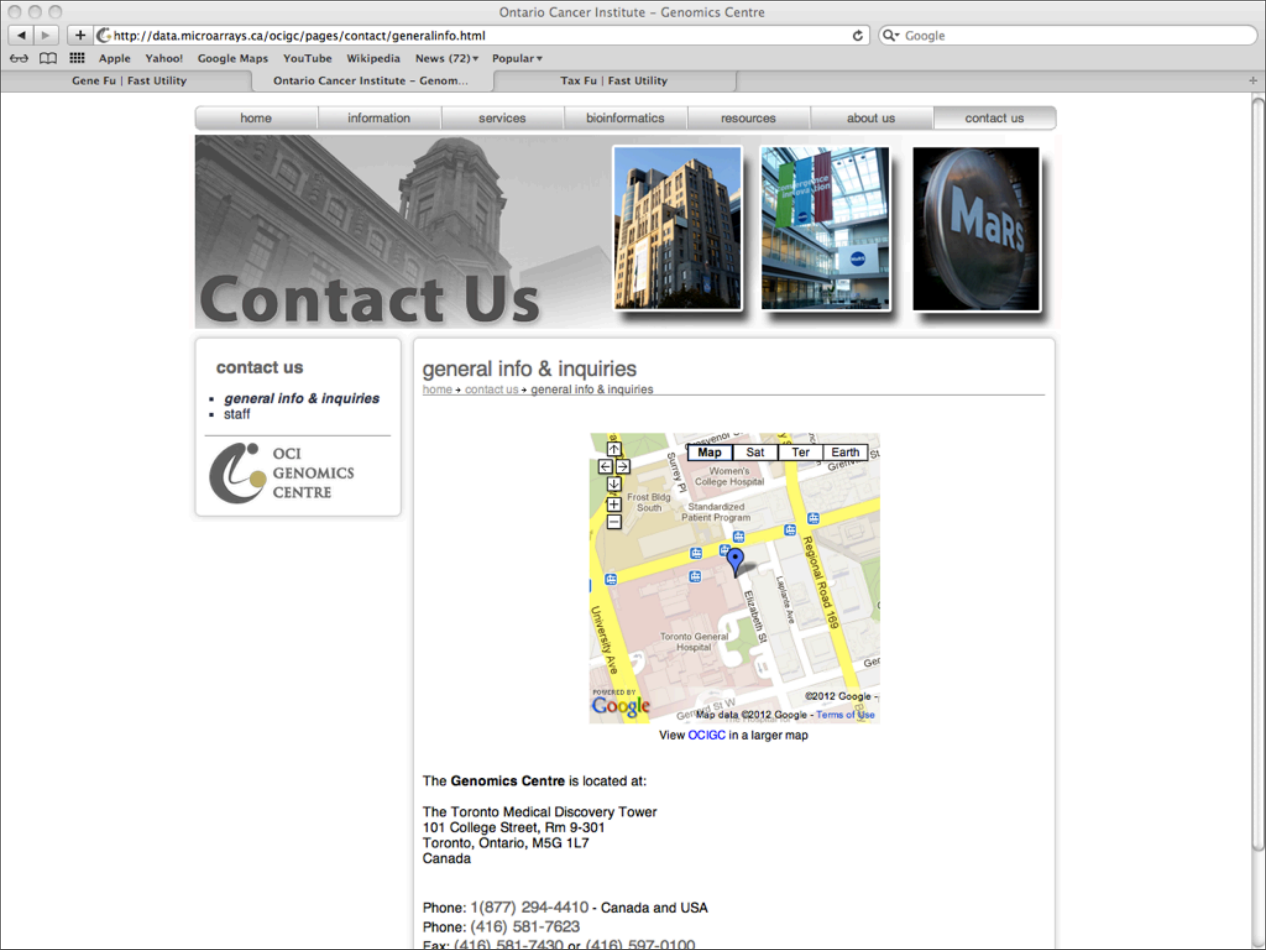
Acknowledgments

OCI/UHN

Qun Jin
Zhibin Lu (formerly, now @ OICR)
Natalie Stickle
Neil Winegarden

UWaterloo Co-op Students

Jimmy He
Connie Li
Di Liu
Zaven Nahapetyan
Pavel Petrenko
Fiona Whelan
Jennifer Wong



contact us

- *general info & inquiries*
- staff

OCI GENOMICS CENTRE

general info & inquiries

home → contact us → general info & inquiries

View OCIGC in a larger map

The **Genomics Centre** is located at:

The Toronto Medical Discovery Tower
101 College Street, Rm 9-301
Toronto, Ontario, M5G 1L7
Canada

Phone: 1(877) 294-4410 - Canada and USA
Phone: (416) 581-7623
Fax: (416) 581-7430 or (416) 597-0100

Tuesday, November 27, 2012