A quest for EDC-related genes in the context of bio-medical literature

Massimo Ubaldi, Laura Soverchia, Roberto Ciccocioppo, Gilberto Mosconi and Alberta Maria Polzonetti

Department of Pharmacological Sciences and Experimental Medicine
University of Camerino (UNICAM)

Dipartimento di Scienze Morfologiche e Biochimiche Comparate
University of Camerino (UNICAM)

Developing a custom-made, low-density, microarray chip to detect the presence of the Endocrine Disrupting Compounds (EDCs) in fish tissues.

The EDCs encompass a vast number of synthetic chemicals and naturally occuring compounds that released in the environment can influence endocrine activity on wild vertebrate species and humans.

The assay we are developing is based on the assessment of **trascriptional signature** of zebrafish (*Danio rerio*), as a animal model, following EDCs exposure.

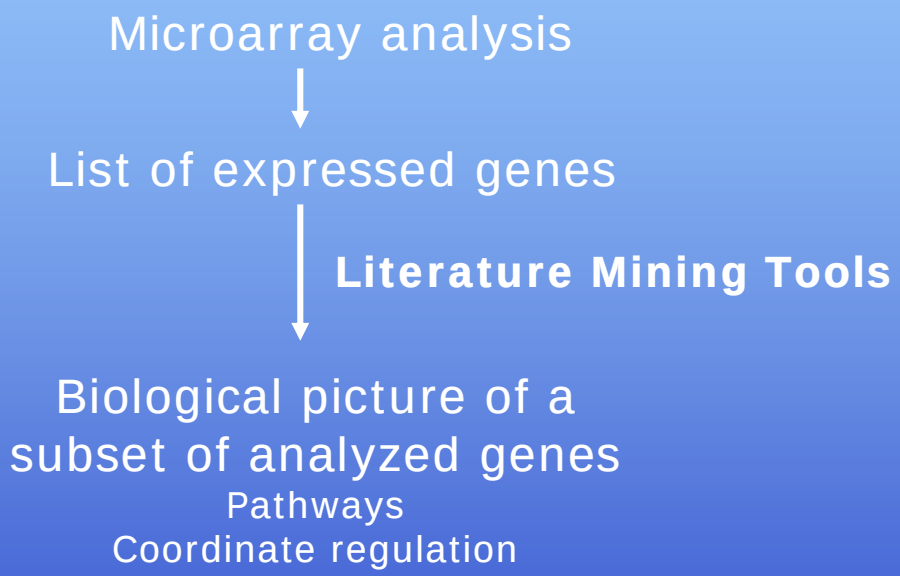
Method

Relevant genes were identified through a knowledge-based search of the existing literature. The literature searches were either of biological databases or journal publications and performed by trained endocrinologists.

GENES UNDER XENOESTROGEN CONTROL		
Cathepsin D (Cat D)	Fish	Carnevali and Maradonna 2003
Eggshell protein (ZR)	Fish	Aruke et al 1997
Oxidase Cytochrome (CYP1A1)	Fish	Anderson et al 1996
Aryl Hydrocarbon Receptor Repressor (AhRR)	Fish	Maradonna et al 2004
Estradiol receptor (ER)	Fish	Polzonetti et al 2004
Vitellogenin (VTG)	Amphibian	Kloas et al 2002
Retinol binding protein (RBP)	Amphibian	Mc Kearin and Shapiro 1988

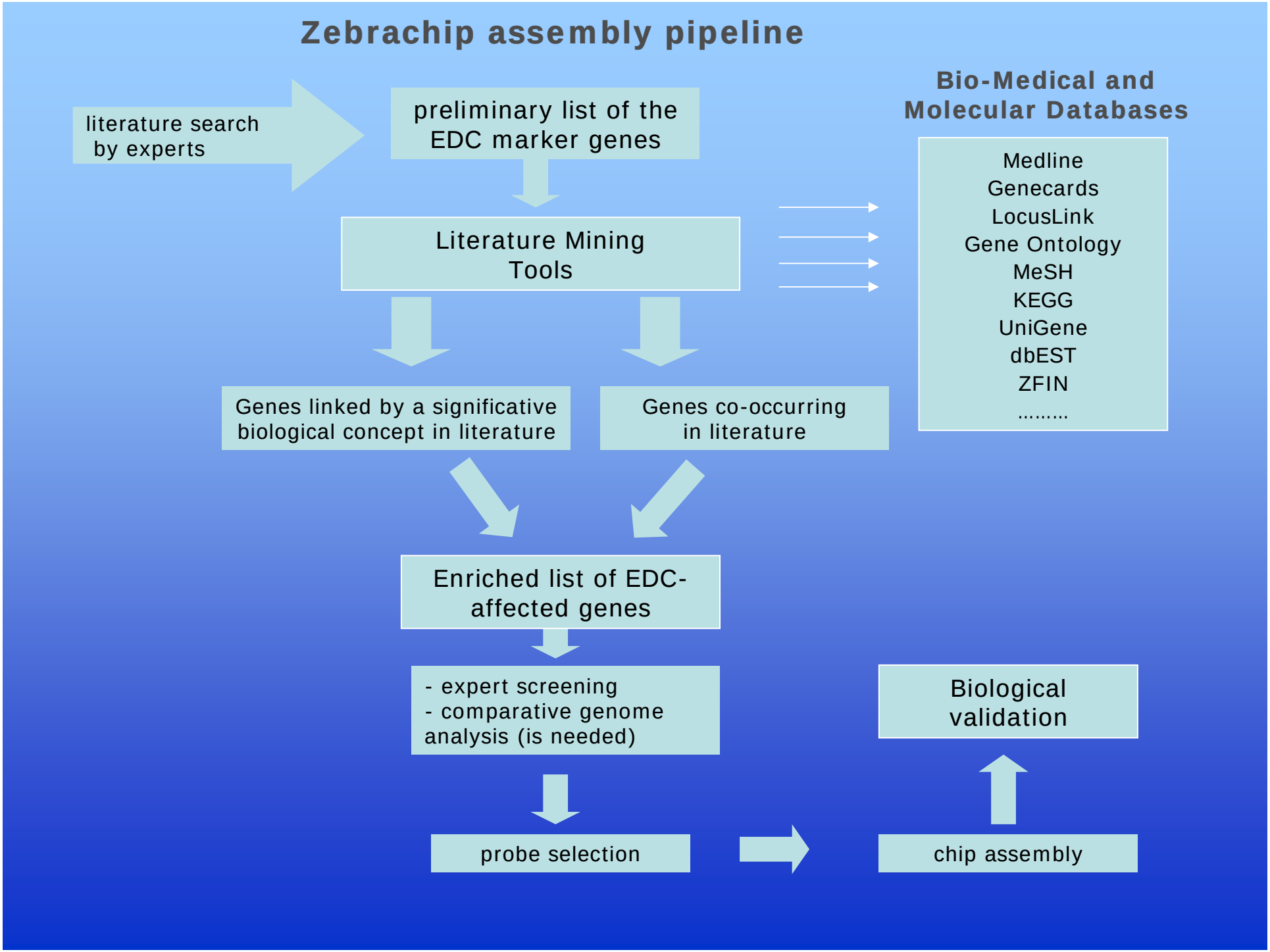
Moreover we utilized some of the available literature mining tools to search for genes potentially related to our biological theme even if not yet directly characterized in that contest.

Some tools are available to analyze the significative expressed genes in the contest of the knowledge collected in the biomedical literature.



Some Available Tools To Mine The Biomedical Literature.

Name	Sources	Description	Usage
MedMiner	GeneCards PubMed	Filters and organizes large amounts of textual and structured information returned from public search engines like GeneCards and PubMed	**Input**: gene names, general concepts, disease names **Output**: relevant citations in the abstracts grouped by keywords
GoMiner	Gene Ontology	Organizes list of genes in the context of the Gene Ontology	**Input**: list of genes **Output**: input genes in the GO tree contest
MedGene	MeSH LocusLink	Summarizes and estimates the realtive strength of all human gene-disease relantionship in Medline.	**Input**: list of genes **Output**: frequency of gene-disease occurrence in Medline
XplorMed	PubMed MeSH	Analyze for relationship between words in a list of abstracts acquired following a PubMed search.	**Input**: keywords **Output**: level of terms association in Medline
PubGene	PubMed GO MeSH	Browse literature or sequence networks, search literature articles and search MeSH or ontology terms associations for a set of genes	**Input**: genes set **Output**: literature, GO, MeSH associations
PubMatrix	PubMed	Simple text based mining of the PubMed	**Input**: genes set, keywords **Output**: frequency of co-occurrence matrix
EASE	LocusLink GenBank UniGene GO PubMed	Biological theme determination for lists of genes; online analysis tool	**Input**: genes set **Output**: annotations, GO categories, gateway to PubMatrix, MedMiner, PubMed, LocusLink, PubGene.
GIS	PubMed	Text-mining system focused on four types of gene-related information: biological functions, associated diseases, related genes and gene–gene relations	**Input**: genes set **Output**: information on biological functions, associated diseases and related genes for the queried list


Zebrachip assembly pipeline
literature search by experts
preliminary list of the EDC marker genes
Literature Mining Tools
Bio-Medical and Molecular Databases
Medline
Genecards
LocusLink
Gene Ontology
MeSH
KEGG
UniGene
dbEST
ZFIN
.....
Genes linked by a significative biological concept in literature
Genes co-occurring in literature
Enriched list of EDC-affected genes
- expert screening
- comparative genome analysis (is needed)
probe selection
chip assembly
Biological validation

PubGene

http://www.pubgene.org/

Jenssen et al., *Nature Genet*ics 28(1):21-8, 2001.

Network of co-occurrence of gene symbols or short term gene names in the title or in abstract in MEDLINE

1. Construction of a gene-article index.
2. Annotation of the gene network with potential biological functions using the keywords or MeSH associated with each paper.
3. Link of two genes if they occurred in the same article.
4. Graphic representation of genes by a node in the network and connecting link if the genes co-occurred.
5. Strength of link: weighted by the number of occurrences of a particular pairing.

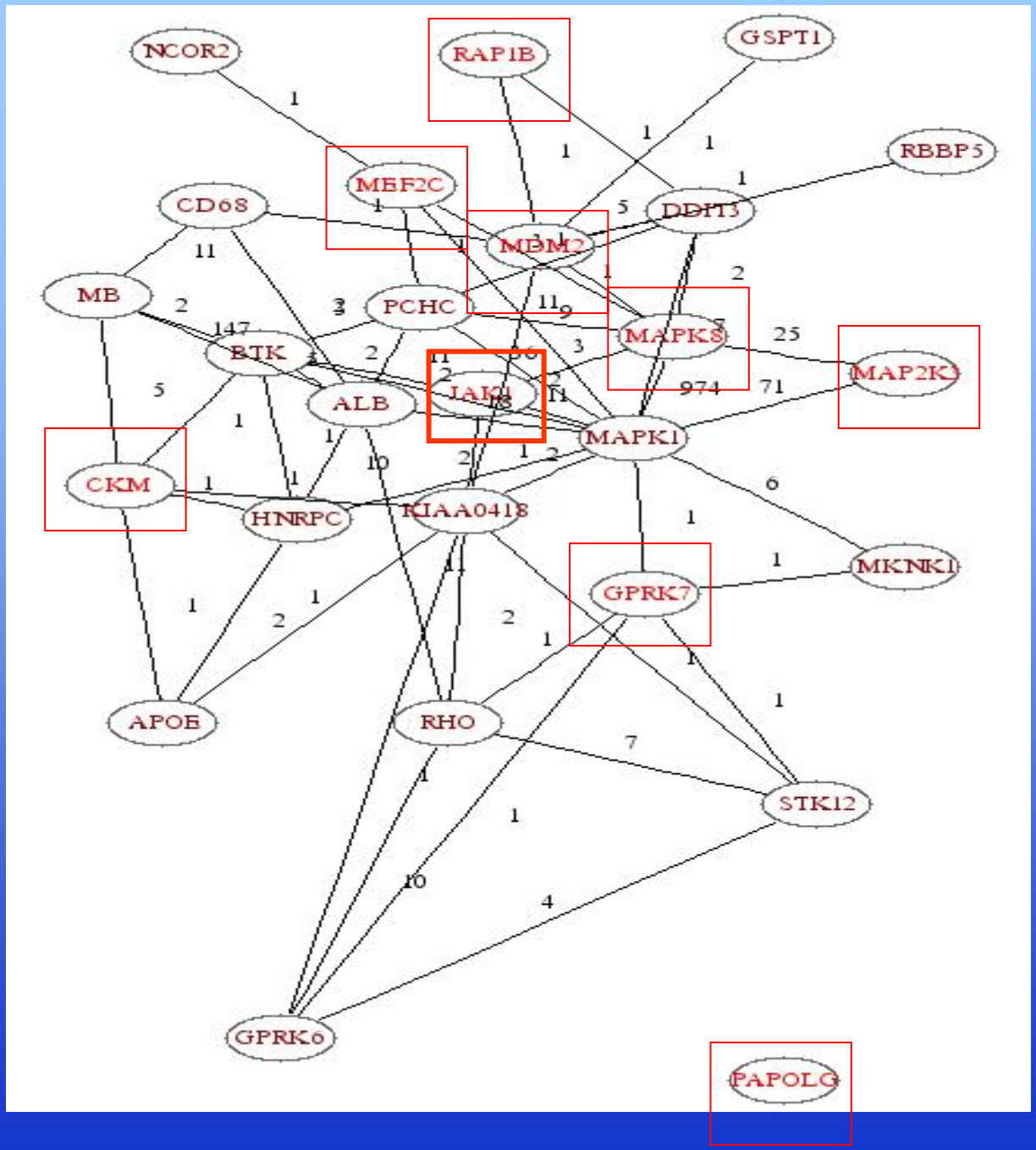
Statistics

(PubGene 2.1, August 2003 update)

	Human	Mouse	Rat
Number of Genes, Primary Symbols	25,528	38,729	5,325
Number of Genes, Synonym Names	28,036	40,636	6,374
Number of Proteins, Primary Symbols	9,172	6,000	3,226
Number of Proteins, Synonym Names	60,413	35,896	20,458
Gene Pair Associations, Text	482,818	49,962	26,056
Protein Pair Associations, Text	1,308,974	842,894	565,382
Gene Pair Associations, Sequence	1,232,545	1,543,828	110,106
Protein Pair Associations, Sequence	276,962	125,486	51,212
MeSH Associations	3,492,050	957,611	669,799
Ontology Associations	643,535	170,590	122,895

Signal Transduction/Cell Cycle
EDC-related genes

Gene	Symbol
mitogen-activated protein kinase kinase 3	map2k3
G protein-coupled receptor kinase 7	gprk7
mitogen-activated protein kinase 8	mapk8
creatine kinase, muscle	ckm
myocyte enhancer factor 2c	mef2c
myogenic differentiation	myod
heat shock protein 90-beta	hsp90b
RAS related protein 1b	rap1b
murine double minute 2 homolog	mdm2
elongation factor 1-alpha	ef1a
poly(A) polymerase gamma	Papolg


NCOR2
RAP1B
GSPT1
RBBP5
MEF2C
CD68
DDIT3
MDM2
MB
PCHC
MAPK8
BTK
MAP2K3
JAK1
ALB
MAPK1
CKM
HNRPC
KIAA0418
MKNK1
GPRK7
APOE
RHO
STK12
GPRK6
PAPOLG

JAK1
Janus kinase 1
(a protein tyrosine kinase)

Locus ID: *Hs* 3716
Dr 30280

Comparative evaluation

Functional evaluation

JAK1 is a large, widely expressed membrane-associated phosphoprotein. JAK1 is involved in the interferon-alpha/beta and -gamma signal transduction pathways.

Homologene analysis

Literature

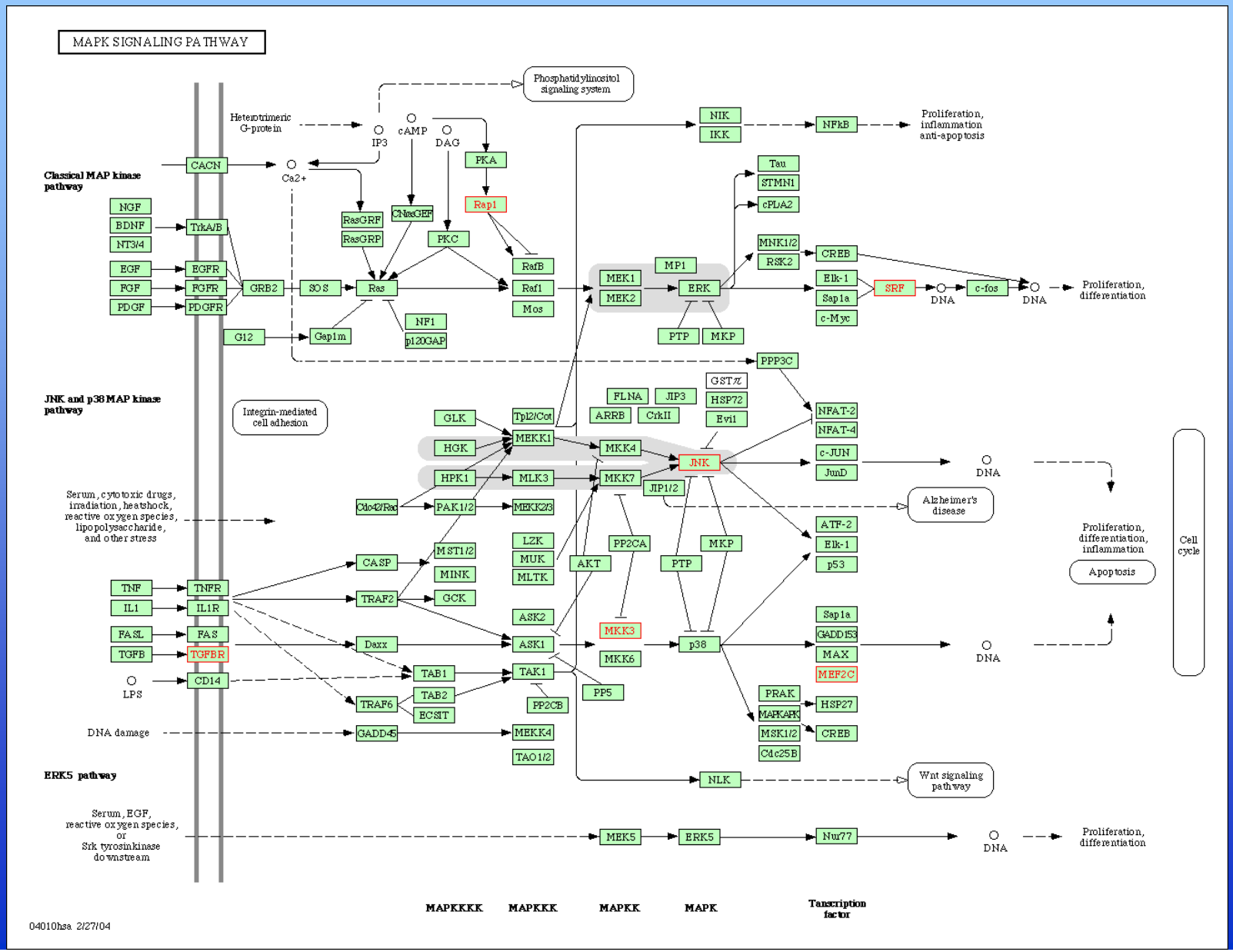
ProtEST
62.93 % / 1137 aa similarity with Hs

Hortologue evidences:
A genetic linkage map for zebrafish: comparative analysis and localization of genes and expressed sequences
Gates, M.A. et al.
Genome Res. 9(4):334-347

Other tools:
BLAST
Pipmaker
Blat

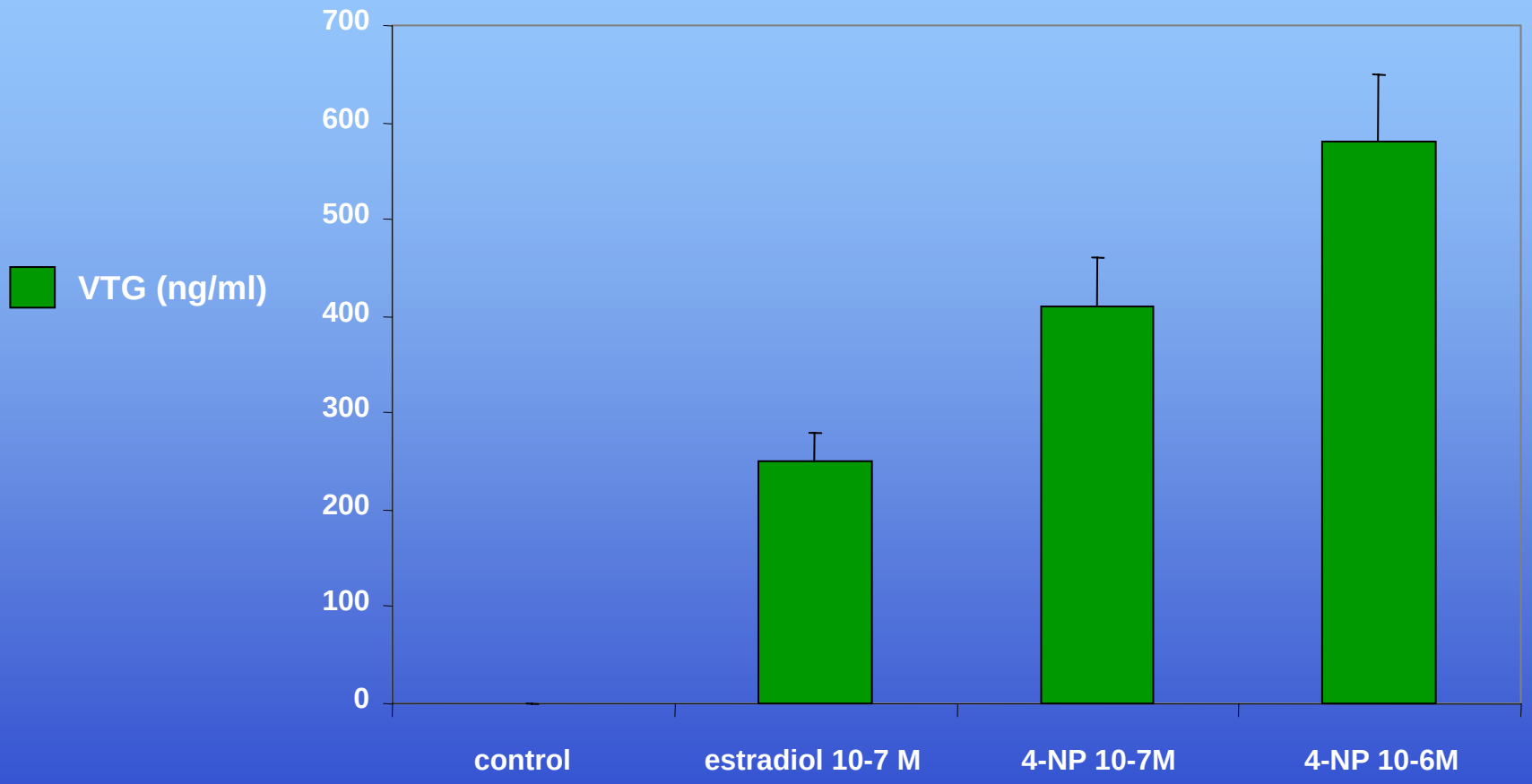
KEGG Pathways Database

(Kyoto Encyclopedia of Genes and Genomes)
http://www.genome.ad.jp/kegg/kegg2.htm



Zebrafish GENE	Symbol	Locus ID
glucose-6-phosphate dehydrogenase	g6pdl	30720
cytochrome P450, subfamily I, polypeptide 1	cyp1a1	140634
cytochrome P450, 19a	cyp19a	30390
glutathione S-transferase M	gstm	324366
peroxisome proliferator activated receptor beta	pparb	30750
fatty acid desaturase 2	fads2	140615
lactate dehydrogenase	ldha	30496
aryl hydrocarbon receptor 1	ahr1	246224
mt cytochrome c oxidase subunit II	mtco2	140540
mt cytochrome c oxidase subunit III	mtco3	140541
mt cytochrome b	mtcyb	140512
hydroxysteroid (17-beta) dehydrogenase	hsd17b4	393105
Vitellogenin 1	vg1	64257
zona pellucida glycoprotein	zp2	30593
retinol binding protein	rbp1	171477
macrophage stimulating 1 (hepatocyte growth factor-like)	mst1	259260
integrin, beta 1	itgb1	378714
RAS related protein 1b	rap1b	215449
murine double minute 2 homolog	mdm2	30637
elongation factor 1-alpha	ef1a	30516
myogenic differentiation	myod	30513
mitogen-activated protein kinase kinase 3	map2k3	65239
G protein-coupled receptor kinase 7	gprk7	373871

Zebrafish GENE	Symbol	Locus ID
heat shock protein 47	hsp47	30449
heat shock protein 70	HSP70	30671
metallothionein	mt	30282
cathepsin D	ctsd	65225
jun B proto-oncogene	JUNB	407086
tumor protein p53	P53	30590
c-fos serum response factor	SRF	30431
parathyroid hormone receptor 1	Pthr1	30629
vitellogenin 3, phosvitinless	vg3	30518
thyroglobulin	tr	368212
matrix metalloproteinase	mmp2	337179
myosin regulatory light chain	mylip	335888
actin, alpha 1	acta1	58114
keratin 8	krt8	352911
gonadotropin-releasing hormone 2	gnrh2	353222
activin A receptor, type IB	acvr1b	30183
activin receptor IIa	acvr2a	30437
inhibin, beta B	inhbb	30275
follistatin	fst	30235
janus kinase	jak1	30280
mitogen-activated protein kinase 8	mapk8	65236
creatine kinase, muscle	ckm	30095
myocyte enhancer factor 2c	mef2c	30572
heat shock protein 90-beta	hsp90b	30573


VTG (ng/ml)
700
600
500
400
300
200
100
0
control
estradiol 10-7 M
4-NP 10-7M
4-NP 10-6M