Trip into the Desert

Drosophila melanogaster & Co

Sven Findeiß

Lehrstuhl für Bioinformatik
University of Leipzig

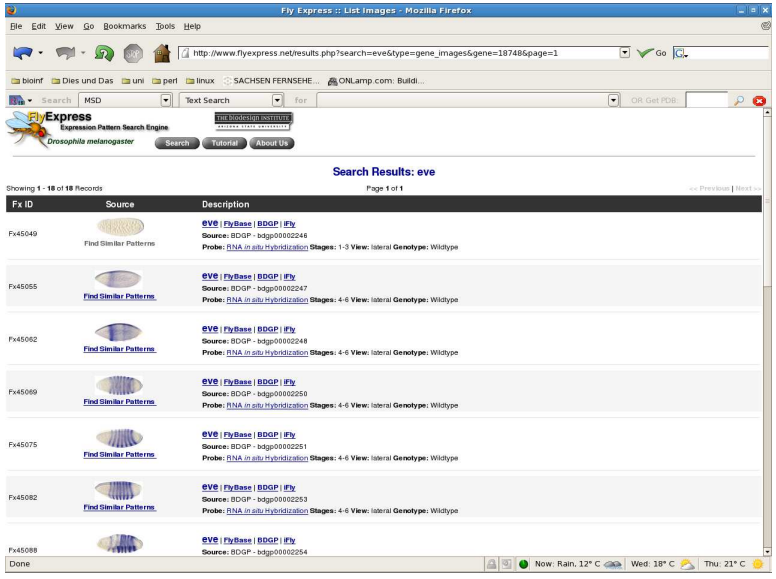
Outline

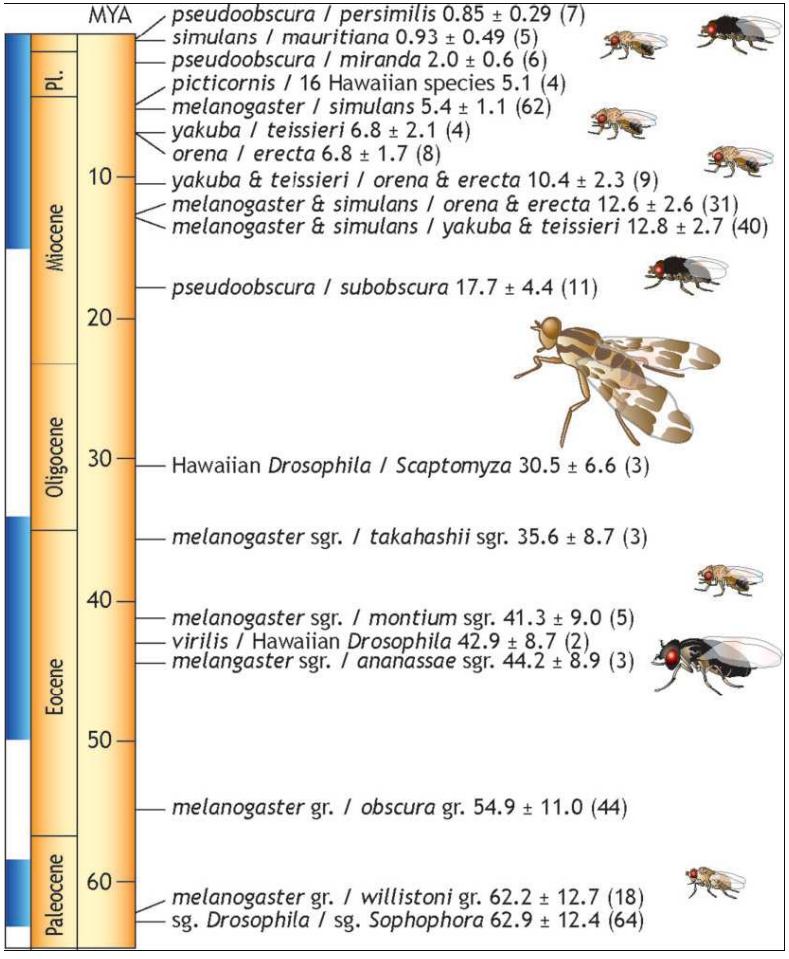
Outline


DREAMSTIME.COM

Center for Evolutionary Functional Genomics (EFG)

Current research achievements:



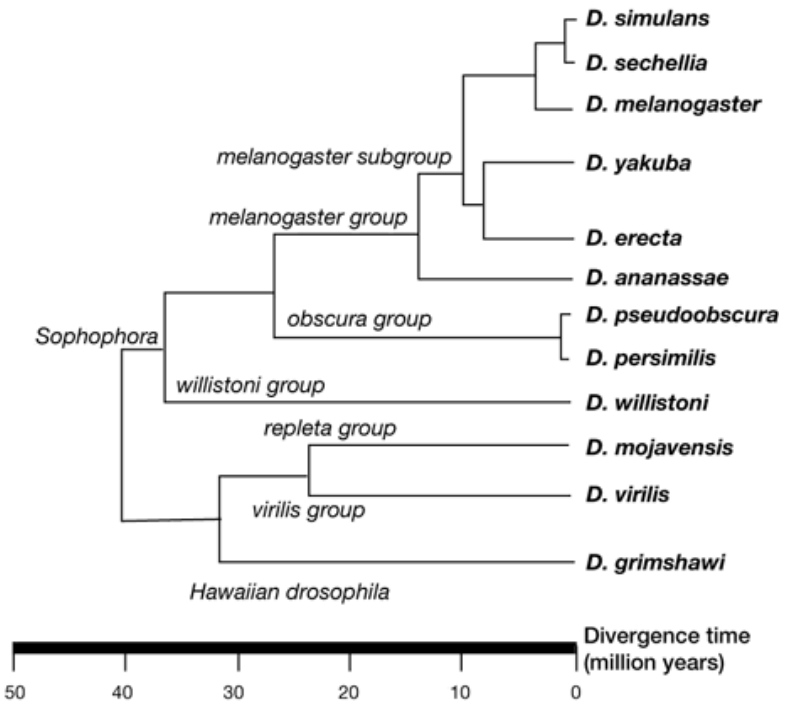


Outline

Motivation

Why analyze *Drosophila* genomes?

- *Drosophila melanogaster* is a well studied model organism
- high homology of *Drosophila melanogaster* and human (584 *Drosophila* genes related to human disease [a])
- 11 recently sequenced non-*melanogaster* species



[a] http://superfly.ucsd.edu/homophila/

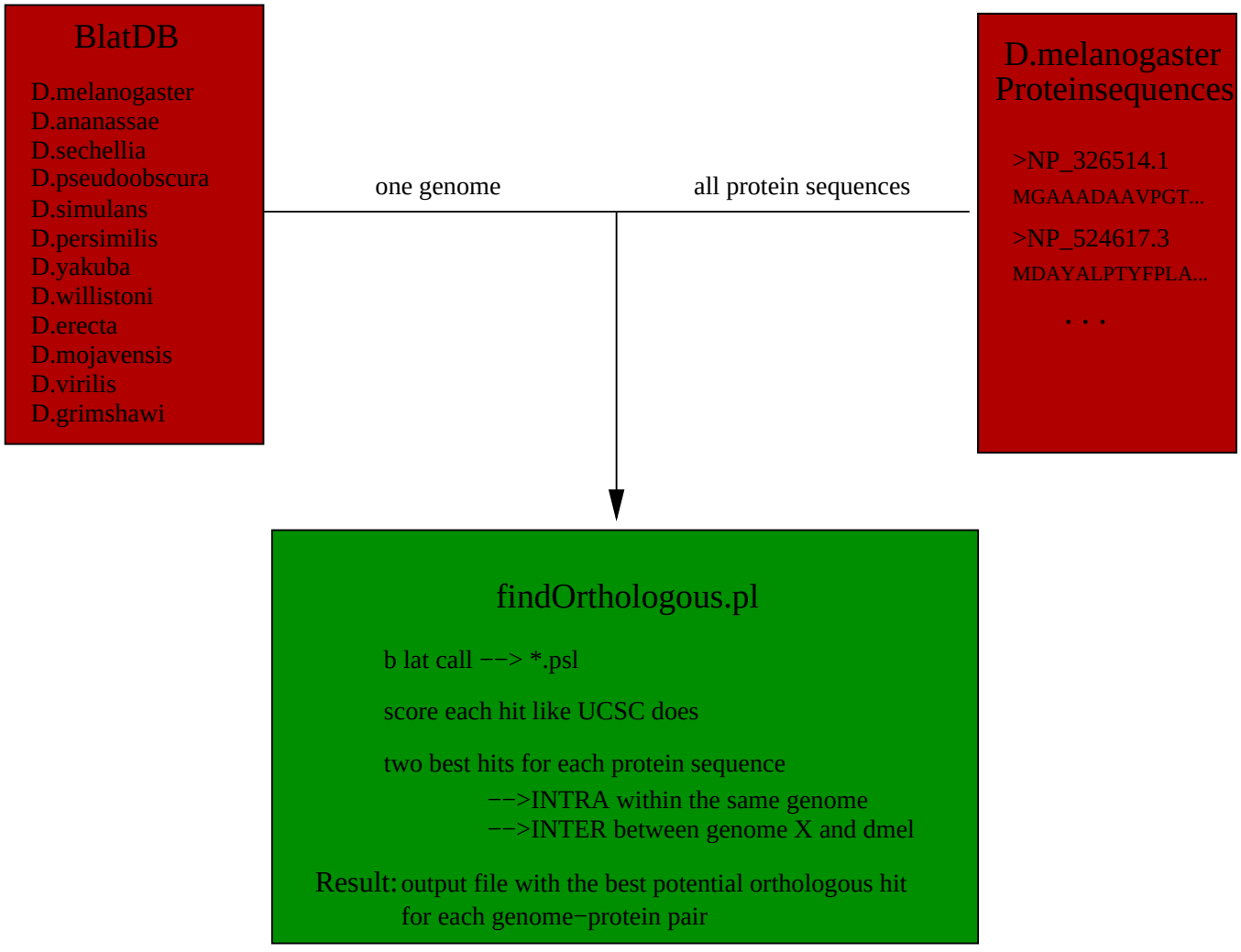
Outline

Idea

Use the blat algorithm to search for orthologous protein sequences in the 11 non-*melanogaster* species.

- *Drosophila melanogaster* is a well studied model organism
  $\Rightarrow$ list of annotated protein sequences downloaded from `http://genome.ucsc.edu`
- *melanogaster* and recently sequenced genomes downloaded from `http://rana.lbl.gov/drosophila/caf1.html`
- *Drosophila melanogaster* was used as reference sequence

Workflow


BlatDB
D.melanogaster
D.ananassae
D.sechellia
D.pseudoobscura
D.simulans
D.persimilis
D.yakuba
D.willistoni
D.erecta
D.mojavensis
D.virilis
D.grimshawi
one genome
all protein sequences
D.melanogaster Proteinsequences
>NP_326514.1
MGAAADAAVPGT...
>NP_524617.3
MDAYALPTYFPLA...
. . .
findOrthologous.pl
b lat call −−> *.psl
score each hit like UCSC does
two best hits for each protein sequence
−−>INTRA within the same genome
−−>INTER between genome X and dmel
Result: output file with the best potential orthologous hit
for each genome−protein pair

Some details

Calculation of Scores

$$UCSC_{score} = 3 \times (match + (repMatch >> 1)) - 3 \times misMatch - qNumInsert - tNumInsert$$

$$INTRA_{score} = \frac{best\ UCSC_{score} - second\ best\ UCSC_{score}}{best\ UCSC_{score}}$$

$\Rightarrow$ value close to 1 shows the uniqueness of a hit

$$INTER_{score} = \frac{best\ UCSC_{score}\ current}{best\ UCSC_{score}\ reference}$$

$\Rightarrow$ value close to 1 means high similarity to reference hit (dmel protein against dmel genome)

The output (1)

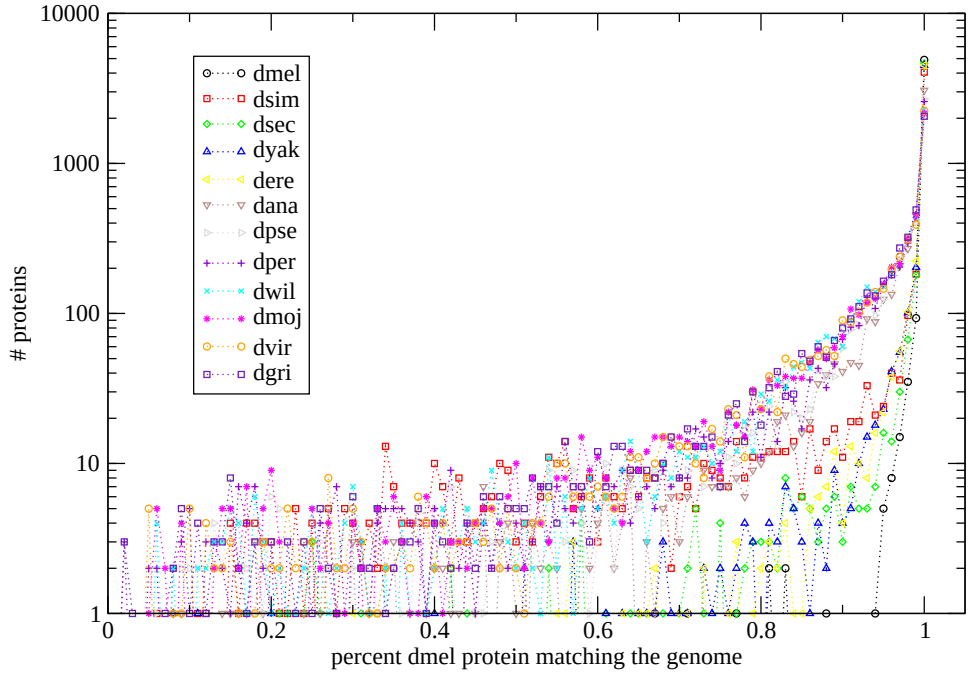
emacs@localhost.localdomain

File Edit Options Buffers Tools Help

#NAME	CHROMOSOM	PROTEIN	STRAND	SCORE	INTRA	INTER	NTERM	CTERM	HITSTART	HITEND
dmel_NP_001007096.1	dmel_3L	NP_001007096.1	-	4557	0.941409	1.000000	0.000000	0.000000	19685705	19692073
dana_NP_001007096.1	dana_scaffold_13337	NP_001007096.1	+	3755	0.934221	0.824007	0.000000	0.000000	17177309	17183397
dere_NP_001007096.1	dere_scaffold_4784	NP_001007096.1	-	4399	0.880200	0.965328	0.000000	0.000000	19506009	19512662
dgri_NP_001007096.1	dgri_scaffold_15110	NP_001007096.1	+	3114	0.885035	0.683344	0.000000	0.000000	16662632	16669501
dmoj_NP_001007096.1	dmoj_scaffold_6680	NP_001007096.1	-	3307	0.888419	0.725697	0.000000	0.000000	12260037	12266315
dper_NP_001007096.1	dper_super_48	NP_001007096.1	-	2743	0.875319	0.601931	0.000000	0.276721	17526	574971
dpse_NP_001007096.1	dpse_ChXR_group8	NP_001007096.1	-	3747	0.916200	0.822251	0.000000	0.004590	7788295	7794126
dsec_NP_001007096.1	dsec_super_29	NP_001007096.1	-	4366	0.948923	0.958086	0.000000	0.000000	411269	417574
dsim_NP_001007096.1	dsim_chr3L	NP_001007096.1	-	4511	0.920195	0.989906	0.000000	0.000000	19039455	19046110
dvir_NP_001007096.1	dvir_scaffold_13049	NP_001007096.1	+	3380	0.900000	0.741716	0.000000	0.000000	18965491	18972621
dwil_NP_001007096.1	dwil_scaffold_180949	NP_001007096.1	-	3345	0.889088	0.734036	0.000000	0.000000	4854927	4861693
dyak_NP_001007096.1	dyak_chr3L	NP_001007096.1	+	4403	0.399727	0.966206	0.000000	0.000000	20807162	20813839
dmel_NP_001007097.1	dmel_3L	NP_001007097.1	-	4557	0.941409	1.000000	0.000000	0.000000	19685705	19692073
dana_NP_001007097.1	dana_scaffold_13337	NP_001007097.1	+	3755	0.934221	0.824007	0.000000	0.000000	17177309	17183397
dere_NP_001007097.1	dere_scaffold_4784	NP_001007097.1	-	4399	0.880200	0.965328	0.000000	0.000000	19506009	19512662
dgri_NP_001007097.1	dgri_scaffold_15110	NP_001007097.1	+	3114	0.885035	0.683344	0.000000	0.000000	16662632	16669501
dmoj_NP_001007097.1	dmoj_scaffold_6680	NP_001007097.1	-	3307	0.888419	0.725697	0.000000	0.000000	12260037	12266315
dper_NP_001007097.1	dper_super_48	NP_001007097.1	-	2743	0.875319	0.601931	0.000000	0.276721	17526	574971
dpse_NP_001007097.1	dpse_ChXR_group8	NP_001007097.1	-	3747	0.916200	0.822251	0.000000	0.004590	7788295	7794126
dsec_NP_001007097.1	dsec_super_29	NP_001007097.1	-	4366	0.948923	0.958086	0.000000	0.000000	411269	417574
dsim_NP_001007097.1	dsim_chr3L	NP_001007097.1	-	4511	0.920195	0.989906	0.000000	0.000000	19039455	19046110
dvir_NP_001007097.1	dvir_scaffold_13049	NP_001007097.1	+	3380	0.900000	0.741716	0.000000	0.000000	18965491	18972621
dwil_NP_001007097.1	dwil_scaffold_180949	NP_001007097.1	-	3345	0.889088	0.734036	0.000000	0.000000	4854927	4861693
dyak_NP_001007097.1	dyak_chr3L	NP_001007097.1	+	4403	0.399727	0.966206	0.000000	0.000000	20807162	20813839
dmel_NP_001014582.1	dmel_3L	NP_001014582.1	-	2006	0.865902	1.000000	0.000000	0.000000	12437218	12441565
dana_NP_001014582.1	dana_scaffold_13337	NP_001014582.1	+	1499	0.723149	0.747258	0.000000	0.001493	1201915	1205972
dere_NP_001014582.1	dere_scaffold_4784	NP_001014582.1	-	1901	0.885324	0.947657	0.000000	0.000000	12452669	12457014
dgri_NP_001014582.1	dgri_scaffold_15110	NP_001014582.1	+	1224	0.804739	0.610169	0.000000	0.019403	5059343	6158280
dmoj_NP_001014582.1	dmoj_scaffold_6680	NP_001014582.1	-	1219	0.812961	0.607677	0.000000	0.000000	6057840	6061524
dper_NP_001014582.1	dper_super_9	NP_001014582.1	-	1529	0.661871	0.762213	0.000000	0.000000	1537584	1541222
dpse_NP_001014582.1	dpse_ChXR_group6	NP_001014582.1	-	1535	0.827362	0.765204	0.000000	0.000000	3242055	3245717
dsec_NP_001014582.1	dsec_super_0	NP_001014582.1	-	1926	0.868640	0.960120	0.000000	0.000000	4640331	4644534
dsim_NP_001014582.1	dsim_chr3L	NP_001014582.1	-	1929	0.881286	0.961615	0.000000	0.000000	11838230	11842561
dvir_NP_001014582.1	dvir_scaffold_13049	NP_001014582.1	+	1213	0.810387	0.604686	0.000000	0.000000	9289088	9292745

--:-- **droMel.output** (Fundamental)--L1--C0--Top----------

Long Line Truncation enabled



Conclusion:

- fraction of dmel genes matching the genomes represent the relationship of flies (exception dsim)
- data quality: most of the dmel proteins match other drosophila genomes with more than 80% of the sequence