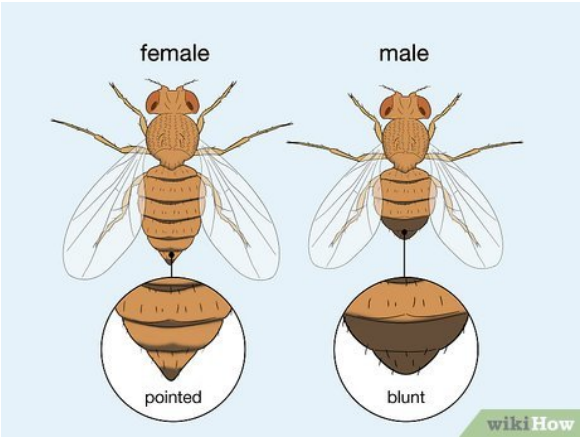
Information theory-based analyses to find a **minimal non-redundant** gene set for sex label classification

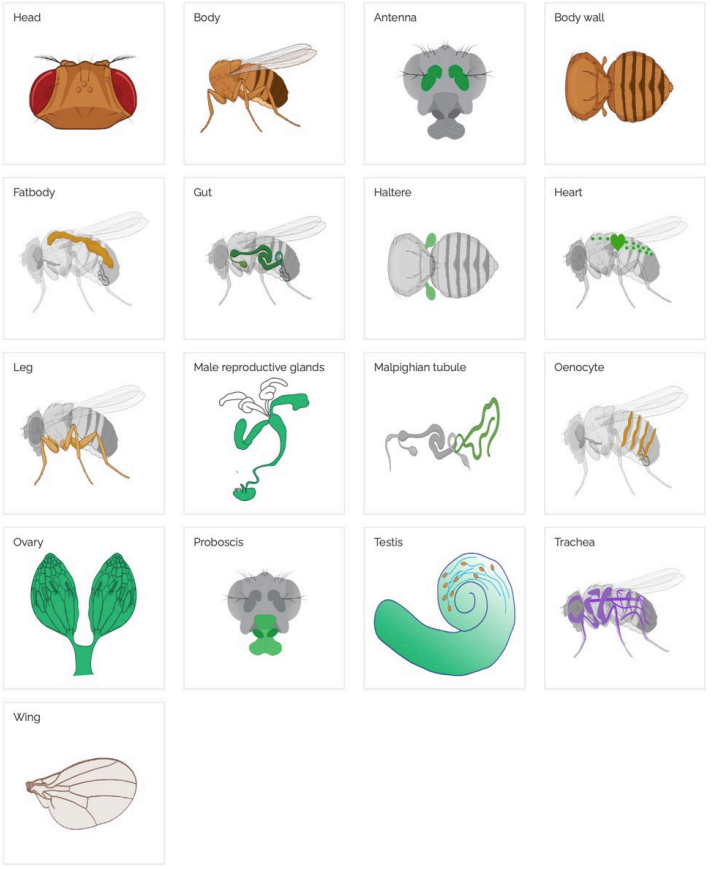
Ming

2021

Question:

- Sex dimorphism at **phenotype** level
- Sex dimorphism at transcription, or **gene expression** level
- Is there a molecular readout of sex dimorphism?
- What are the biological processes/gene expression programs associated with gene expression difference between the two sexes?
- In somatic tissues, does every cell have a ‘sex identity’?
- Are sex identities of different types of cells mediated/manifested by the same set of genes?





Is the sex differentiation potential be realized?

Turn a biological question into a computational one:

- If there is a gene set most informative about the distinct between female and male cells, does this gene set consistent across diverse cell types in a fly?

How to find a minimal non-redundant gene set to predict the sex label of every individual cell

sex labels unknown

	Cell 1	Cell2	Cell3	Cell4	...
Gene 1	0	0	3	10	
Gene 2	24	0	41	12	
Gene 3	175	284	93	162	
Gene 4	0	0	0	0	
Gene 5	36	0	32	21	
...	...	...	...	...	

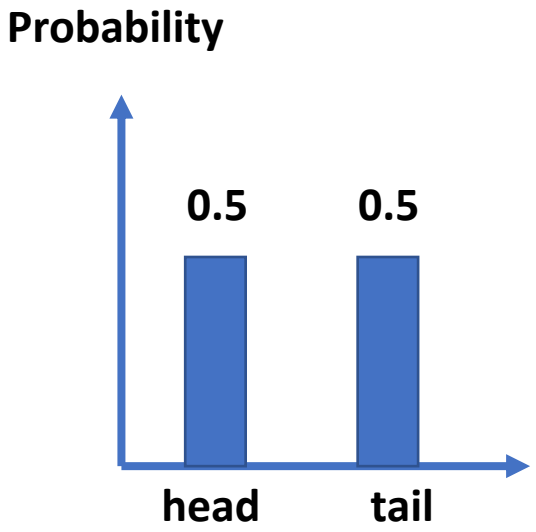
Gene set

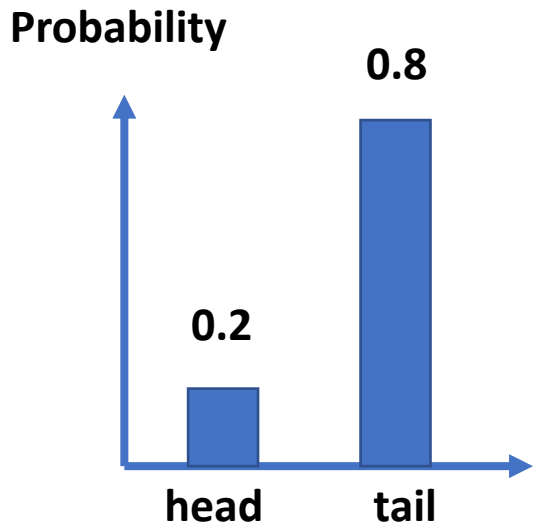
Female male female male

	Cell 1	Cell2	Cell3	Cell4	...
Gene 1	0	0	3	10	
Gene 2	24	0	41	12	
Gene 3	175	284	93	162	
Gene 4	0	0	0	0	
Gene 5	36	0	32	21	
...	...	...	...	...	

Background

Flip a coin, guess head or tail
which one is more uncertain?

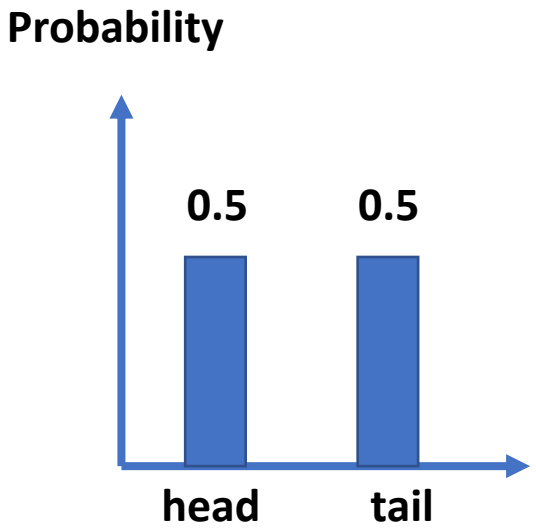




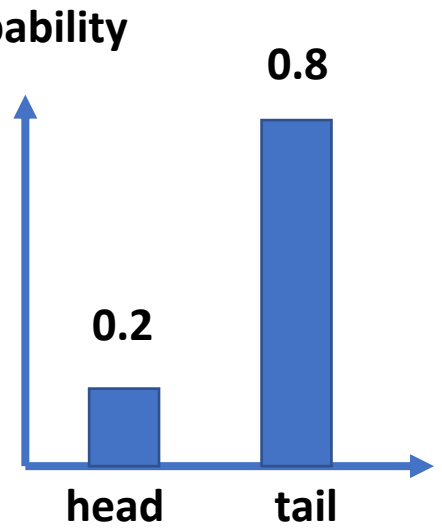
Background

- **Information entropy, a measure of uncertainty**

$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log_b \mathrm{P}(x_i)$$



Flip a coin, guess head or tail
which one is more uncertain?



-1 * {0.5*log(0.5) + 0.5 * log(0.5)} = **0.6931**

-1 * {0.2*log(0.2) + 0.8 * log(0.8)} = **0.5004**

Background

- **Information entropy**, the uncertainty of a variable's possible outcomes.

Larger value ~ higher uncertainty

$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log_b \mathrm{P}(x_i)$$

- **Conditional entropy**, given a second variable (X), what's the amount of the uncertainty still existing in the first variable (Y).

$$\mathrm{H}(Y|X=x) = -\sum_{y \in \mathcal{Y}} \Pr(Y=y|X=x) \log_2 \Pr(Y=y|X=x).$$

Y	1	1	1	1	1	0	0	0	0	0

$H(Y) = -1 * \{ 0.5 * \log(0.5) + 0.5 * \log(0.5) \} = 0.6931$

Y	1	1	1	1	1	0	0	0	0	0
X	1	1	1	0	0	0	0	0	0	1

$H(Y|X) = Pr(X=0) * H(Y|X=0) + Pr(X=1) * H(Y|X=1)$
$= 0.6068$

Mutual information
How much uncertainty about Y decreases when we observe X.

Or the amount of information **gained** knowing X to predict Y

$I(Y;X) = H(Y) - H(Y|X) = 0.6931 - 0.6068 = 0.0863$

Sex label										
	1	1	1	1	1	0	0	0	0	0

$H(Y) = -1 * \{ 0.5 * \log(0.5) + 0.5*\log(0.5) \} = 0.6931$

Sex label	1	1	1	1	1	0	0	0	0	0
Binary gene expression	1	1	1	0	0	0	0	0	0	1

$H(Y|X) = Pr(X=0) * H(Y|X=0) + Pr(X=1)* H(Y|X=1)$
$= 0.6068$

Mutual information

How much uncertainty about Y decreases when we observe X.

Or the amount of information **gained** knowing X to predict Y

$I(Y;X) = H(Y) – H(Y|X) = 0.6931 - 0.6068 = 0.0863$

Algorithm

H(.) information entropy, or uncertainty
I(.) mutual information, or information gained given knowing something

Sex labels
H(S) = 0.8

I (S|Gi)	
gene1	0.3
gene2	0.2
gene3	0.1
gene4	0

Gene set to be filled

Sex labels
H(S|G1) = 0.5

I (S|G1,Gi)	
gene3	0.2
gene4	0.05
gene2	0.02

gene1

Gene set to be filled

Sex labels
H(S|G1,G3) = 0.3

I (S|G1,G3,Gi)	
gene2	0.25
gene4	0.1

gene1
gene3

Gene set to be filled

Sex labels
H(S|G1,G3,G2) = 0.05

I (S|G1,G3,G2,Gi)	
gene4	0.15

gene1
gene3
gene2

Gene set to be filled

Happy with the 'left-over' uncertainty?

Stop the algorithm

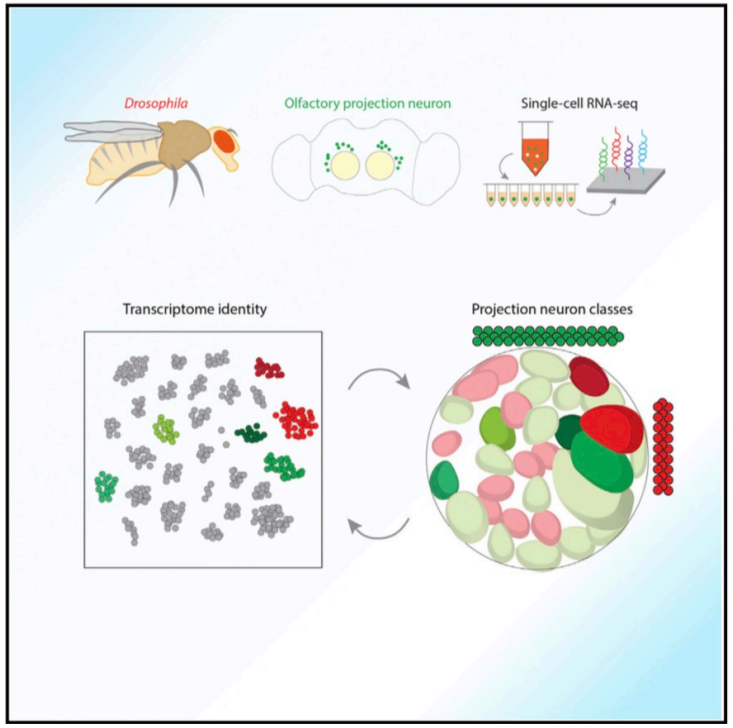
- This algorithm was proposed in this following paper to classify neuron types
- I implemented it in R to derive a gene signature of sex labels of single cells

Resource

Cell

Classifying *Drosophila* Olfactory Projection Neuron Subtypes by Single-Cell RNA Sequencing

Graphical Abstract



Authors

Hongjie Li, Felix Horns, Bing Wu, ..., David J. Luginbuhl, Stephen R. Quake, Liqun Luo

Correspondence

quake@stanford.edu (S.R.Q.), lluo@stanford.edu (L.L.)

In Brief

Single-cell RNA sequencing establishes that transcriptomic identity of *Drosophila* olfactory projection neurons corresponds with connectivity and function and identifies transcription factors and cell-surface molecules as highly informative in encoding neuronal identity.

Dataset

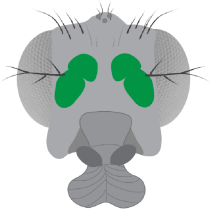
- For each cell type in a tissue, I took the top **100** most informative genes as the input gene list, to initiate the ‘minimal non-redundant gene set’ searching journey.

- Cell type considered: >=20 cells in female and >=20 cells in male

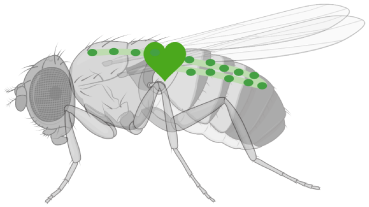
- For each cell type in a tissue, if I want a gene set that could explain 0.99 uncertainty in the classification of sex labels, the resulting gene set would have a certain size, for example, 2 genes, 3 genes, etc.

head

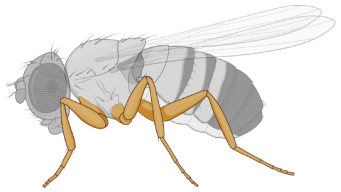
antenna



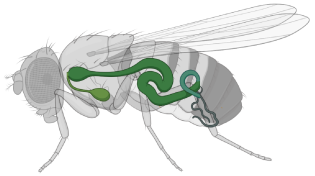
heart



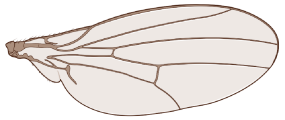
leg



gut



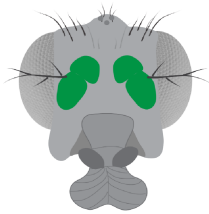
wing

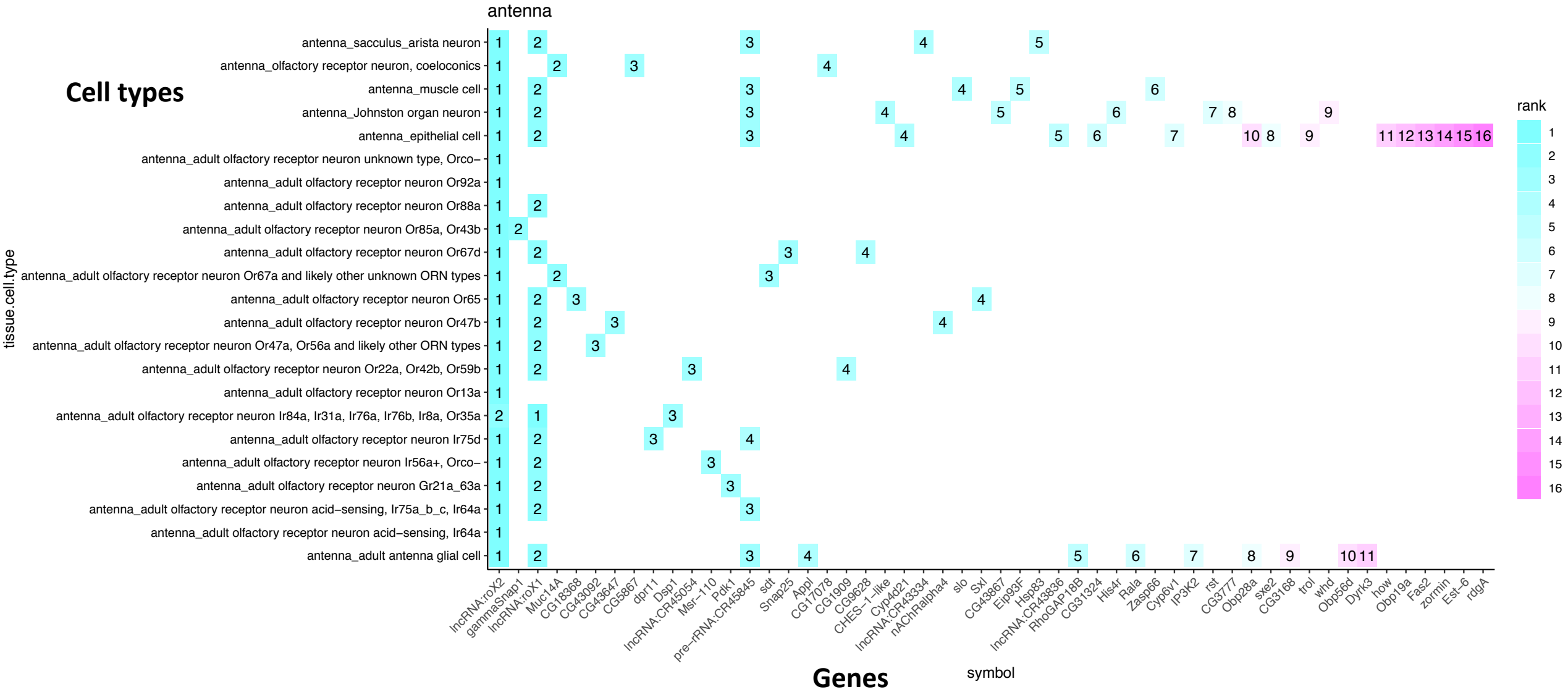


Result

A gene set that could explain **0.99** uncertainty in the classification of sex labels for each cell type

23 cell types in the fly antenna

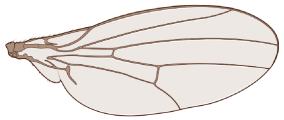


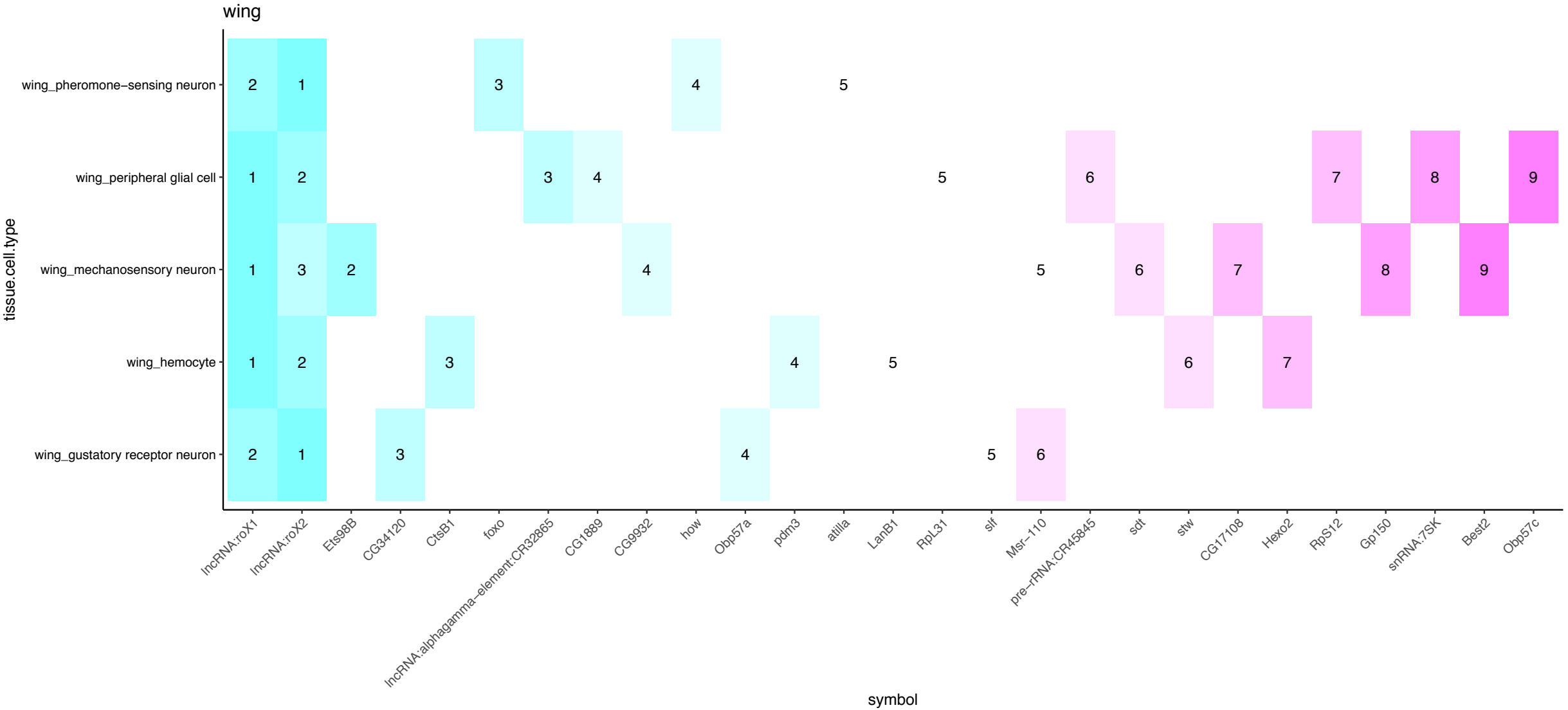


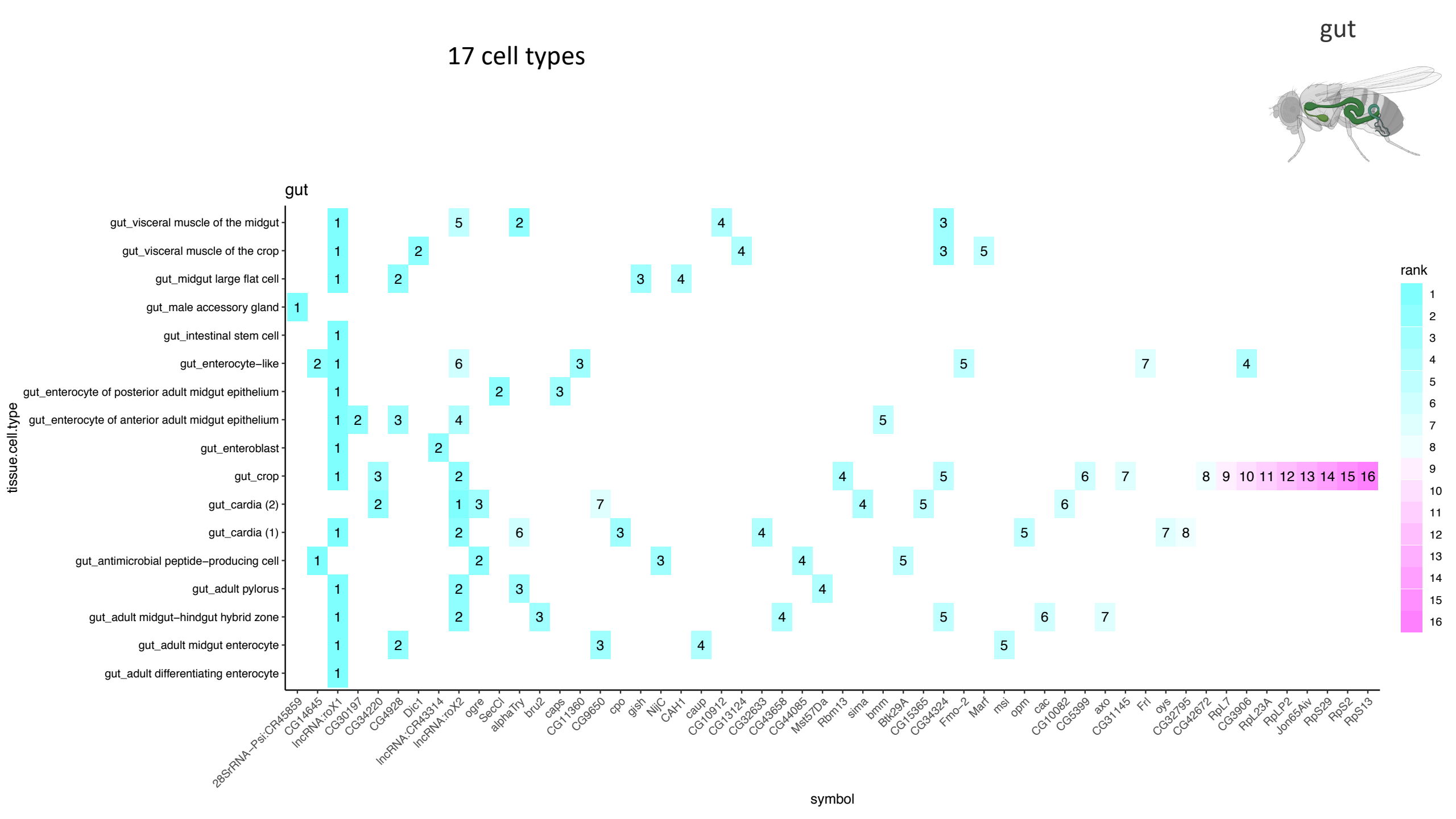
wing

5 cell types

	female	male
Muscle cells	19	125
nociceptive neuron	9	30

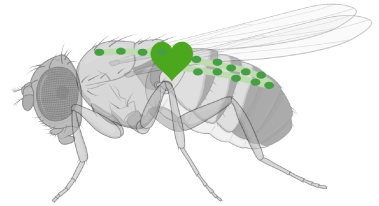





gut
17 cell types
gut
gut_visceral muscle of the midgut
gut_visceral muscle of the crop
gut_midgut large flat cell
gut_male accessory gland
gut_intestinal stem cell
gut_enterocyte−like
gut_enterocyte of posterior adult midgut epithelium
gut_enterocyte of anterior adult midgut epithelium
gut_enteroblast
gut_crop
gut_cardia (2)
gut_cardia (1)
gut_antimicrobial peptide−producing cell
gut_adult pylorus
gut_adult midgut−hindgut hybrid zone
gut_adult midgut enterocyte
gut_adult differentiating enterocyte
tissue.cell.type
28SrRNA−Psi:CR45859
CG14645
lncRNA:roX1
CG30197
CG34220
CG4928
Dic1
lncRNA:CR43314
lncRNA:roX2
ogre
SecCl
alphaTry
bru2
caps
CG11360
CG9650
cpo
gish
NijC
CAH1
caup
CG10912
CG13124
CG32633
CG43658
CG44085
Mst57Da
Rbm13
sima
bmm
Btk29A
CG15365
CG34324
Fmo−2
Marf
msi
opm
cac
CG10082
CG5399
axo
CG31145
Fri
oys
CG32795
CG42672
RpL7
CG3906
RpL23A
RpLP2
Jon65Aiv
RpS29
RpS2
RpS13
symbol
rank
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

12 cell types

heart



heart

heart_muscle cell
heart_indirect flight muscle
heart_hemocyte
heart_epithelial cell
heart_crystal cell
heart_cardiomyocyte, working adult heart muscle (non−ostia)
heart_adult tracheal cell
heart_adult peripheral neuron of the heart
heart_adult oenocyte
heart_adult heart ventral longitudinal muscle
heart_adult fat body
heart_adult abdominal pericardial cell

tissue.cell.type

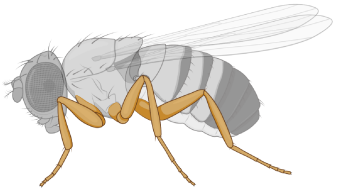
	lncRNA:roX1	Yp1	Nplp3	whe	Yp2	Yp3	Zasp66	Plc21C
heart_muscle cell	1	2					3	
heart_indirect flight muscle	1	2						
heart_hemocyte	2	1			3			
heart_epithelial cell	2	1			3			
heart_crystal cell	2	1						
heart_cardiomyocyte, working adult heart muscle (non−ostia)	1	2						
heart_adult tracheal cell		1			2			
heart_adult peripheral neuron of the heart	1	2	3					
heart_adult oenocyte	1					2		
heart_adult heart ventral longitudinal muscle	2	1				3		
heart_adult fat body	2	1			3			
heart_adult abdominal pericardial cell	1	2		3				4

lncRNA:roX1 Yp1 Nplp3 whe Yp2 Yp3 Zasp66 Plc21C

symbol

rank
1
2
3
4

leg



9 cell types

leg

leg_sensory neuron
leg_peripheral glial cell
leg_muscle cell
leg_mechanosensory neuron of leg chordotonal organ
leg_mechanosensory neuron
leg_hemocyte
leg_gustatory receptor neuron
leg_epithelial cell
leg_adult tracheal cell

tissue.cell.type

lncRNA:roX1
lncRNA:roX2
sls
CG2256
cher
Cyp4d21
jeb
rudhira
sgg
Smox
Argk
Cp1
lncRNA:CR43836
Rbm13
Rgk3
CG42673
crb
Pkc98E
pre-rRNA:CR45845
RhoGAP19D
Cyp313a1
Sema5c
Tet
CG42808
dsx
mei-P26
rdgA
CG33110
kdn
ATPsynF
RhoGAP18B
CG32694
shot
ATPsynG
CG33970
sesB
Xrp1
UQCR-Q

symbol

rank
1
2
3
4
5
6
7
8
9
10
11
12
13
14

head

72 cell types