

### Dm6 dating report 3

The follow content includes comparison between dm3 and dm6 dating result and explanations, mainly concentrate in comparison between dm3 and dm6.

## 1. Gene number comparison

The overall gene number comparison

Table1 gene number comparison between different database versions

	dm3_ver54	dm3_ver69	dm6_flybase	dm6_ensembl(78)
0	11907	11912	12061	12013
1	220	208	201	202
2	161	193	214	214
3	154	200	200	200
4	283	332	334	334
5	68	72	76	76
6	60	41	44	44
total	<b>12853</b>	<b>12958</b>	<b>13080</b>	<b>13083</b>

## 2. Filtered gene number in ver78

Table 2 filtered process gene number

operation	Reduce num	Retain num	table
start	0	13918	Axt_branch
Filter note! ="NA" or pep_seq=""	7	13911	axt_branch_final
Filter genes of chrY and M	35	13876	axt_branch_gene_1
Filter pre, post pair	<b>727</b>	13149	axt_branch_gene_2
Histone gene	54	13095	axt_branch_gene
Ste: gene	12	13083	

### 2.1 Supplementary information of table 2

Fig 1 filter gene number according to (pre, post) value

branch	pre	post	count(*)
0	2	0	355
1	1	0	58
1	2	0	23
2	2	1	27
2	3	1	90
3	1	0	10
3	2	0	24
4	2	0	7
4	2	1	5
4	3	1	12
4	4	1	41
5	1	1	18
5	2	1	28
6	0	1	29

Histone gene example: FBgn0053894, FBgn0053810, FBgn0053827

Ste: gene: FBgn0053238, FBgn0053239, FBgn0053810(CG33810), FBgn0053245

Some ste gene don't location on subterminal region(Seminal fluid protein 23F (FBgn0259949) FBgn0259951).

Fig 2 different dm version subterminal region

	release 4	release 5	size Rel 5	release6	chrom_len
X	20.369.021	20429372	22,422,827	chrX:20558345-23020991	
2L	19.669.505	19673576	23,011,544	chr2L:19673576-23112624	23513712
2R	2.692.485	0-3068177	21,146,708	chr2R:4112496-7181672	25286936
3L	20.545.022	20604281	24,543,557	chr3L:20611181-24550457	28110227
3R	2.811.816	0-2811816	27,905,053	chr3R:4174279-6986094	32079331
4 whole chr		whole chr	1,351,857	chr4:3428-1348131	1348131

### 3 Detailed gene number comparison and explanation

#### 3.1 comparison and explanation between dm6\_ver78 (FB\_2014\_5) and dm3\_flybase (FB\_2014\_6)

Table 3 detailed comparison between dm6\_ver78 and dm3\_flybase

dm6_ver78(FB2014_6) VS dm6_flybase(FB_2014_6)												
		dm6_flybase								total	ver78_total	ver78_only
ver78	0	12001	0	0	0	0	0	0	0	12001	12013	12
	1	1	201	0	0	0	0	0	0	202	202	0
	2	0	0	213	0	0	0	0	0	213	214	1
	3	0	0	0	200	0	0	0	0	200	200	0
	4	0	0	0	0	334	0	0	0	334	334	0
	5	0	0	0	0	0	76	0	0	76	76	0
	6	0	0	0	0	0	0	44	0	44	44	0
	total	12002	201	213	200	334	76	44	44	13070	13083	13
	flybase_total	12011	201	214	200	334	76	44	44	13080		
	flybae_only	9	0	1	0	0	0	0	0	10		

young total identity: (201+213+200+334+76+44)/(13070-12002)=1

These versions almost have identity result. There is few differences due to there are belong different gene annotation data. Ensemble metazoa dm6 annotation data produced in May 2014(flybase r6.02) while my dating annotation data produced in June 2014(flybase r6.03).

The follow table will present detail information about these different genes.

Table 4 Different gene detailed information

Ver78_only	Flybase_only	reason
FBgn0000413->	FBgn0267821	Just change geneID
FBgn0001332->	FBgn0267824	Just change geneID
FBgn0023423->	FBgn0267841	Just change geneID
FBgn0034794->	FBgn0267823	Just change geneID
FBgn0058196->	FBgn0267861	Just change geneID
FBgn0086377->	FBgn0267849	Just change geneID
FBgn0030852->	FBgn0267967	From a unknown coding gene change to a corolla coding and geneID change, but other information didn't change
FBgn0036013->	FBgn0267796	Position changed and transcript number from 1 added to 2
FBgn0036017->	FBgn0267796	FBgn0036013 and FBgn0036017 combined to a gene
FBgn0021953->	FBgn0267828	Just change geneID
FBgn0266458->	FBgn0267828	From unknown coding gene changed to Fatty acid (long chain) transport protein coding gene and transcript number from 2 added to 4, FBgn0021953 and FBgn0266458 combined to a geneID
	FBgn0267964	New geneID in dm6_flybase (FB_2014_6)
FBgn0030758		From Calcineurin A at 14 coding gene change to unknowgene ncRNA(FBgn0267912) and geneID changed but other information didn't change
FBgn0261451		From terribly reduced optic lobes coding gene changed to unknown ncRNA (FBgn0267911)and changed geneID

Num:13	Num:10	
<b>A another exception gene: FBgn0032969</b> branch changed from ensemble 78 (branch: 1) to flybase (branch:0) Reason: this gene only have a transcript (FBtr0344210) in ensemble78(2024-5), but there have two transcripts(FBtr0347503 and FBtr0347504), one transcript longer than previous, so this gene has an older age		

### 3.2 comparison and explanation between dm6\_ver78 (FB\_2014\_5) and dm3\_ver69

Table 5 detailed comparison between dm6\_ver78 and dm3\_ver69

		dm6_ver78(FB_2014_5) VS dm3_ver69									
		dm3_ver69									
		0	1	2	3	4	5	6	total	ver78_total	ver78_only
ver78	0	11364	10	3	12	13	0	0	11402	12013	611
	1	2	174	5	1	3	0	0	185	202	17
	2	0	11	161	5	1	0	0	178	214	36
	3	2	0	3	169	4	0	0	178	200	22
	4	2	0	0	1	290	0	0	293	334	41
	5	0	0	0	0	2	62	2	66	76	10
	6	0	0	0	0	0	0	31	31	44	13
	total	11370	195	172	188	313	62	33	12333	13083	750
	ver69 total	11912	208	193	200	332	72	41	12958		
	ver69 only	542	13	21	12	19	10	8	625		
young gene identity: (174+161+169+290+62+31) / (12333-11370)=0.92											

Note: dm6\_ver78 and dm3\_ver69 have high identity.

The follow table may explain major reason of the differences.

Table 6 Different gene detailed information of these two versions

Explanation of only exist in ver78 or ver69 gene		
	num	reasons
Ver78 only gene:761 Other 13 gene I can't find reason	556	Ver78 has 556 new coding geneID never exit in ver69(such as FBgn0267728, FBgn0267790) Maybe from ver69 geneID changed
	8	Gene biotype changed, ver78 coding gene is ver69 other type gene(pseudogene:5,ncRNA:2) Such as :pseu: FBgn0026399, FBgn0031726, ncRNA: FBgn0047000,
	133	Due to gene (pre.post) filtered in ver69 so only in ver78 exit
	5	These gene in chrU in ver69, but in chr2R or chrX in ver78 such as : FBgn0039993, FBgn0085521
	46	These gene in axt_branch table of ver69 note= lowChr and pep_seq!= ' ' but in ver78 version their note=NA, such as FBgn0259821, FBgn0260008, FBgn0085736
Ver69 only gene:624 : Other 13 gene I can't find reason	468	491 coding geneID in ver69 discarded or changed geneID in ver78 (such as FBgn0262947, FBgn0263255) changed to FBgn0263395)
	75	Gene type from ver69 coding changed to ver78 lincRNA (62(branch:0: 48,4:9)or pseudogene(13(branch:6:6)) Such as FBgn0262033, FBgn0262142, FBgn0262806, FBgn0262546
	81	Due to gene (pre.post) filtered in ver78 so only in ver69 exit
Difference explanation between ver78 and ver54		
version	Branch change	reasons
Ver69->ver78	0->2	Shorter transcript (FBgn0039428, FBgn0042129)
	1->0	More transcript and longer transscript(FBgn0030247, FBgn0033404)
	1->2	Transcript position changed (FBgn0010294, FBgn0053702) or maybe new genome assembly (FBgn0054051 no changed)
	2->1	More transcript and longer transcript(FBgn0003060, FBgn0038741) or maybe new assembly

		(FBgn0262811 no changed)
	2->3	New genome assembly (FBgn0035496) or shorter transcript(FBgn0262583, FBgn0038123)
	3->0	Longer transcript and more transcript (FBgn0032110, FBgn0262030, FBgn0085236)
	3->4	Maybe new genome assembly (FBgn0003683 other information no change)
	4->0	Longer transcript and more transcript (FBgn0031331, FBgn0261584)
	4->5	Shorter transcript and more transcript(FBgn0028986) or new genome assembly(FBgn0030937)
<b>Note:</b> in general, from ver69 to ver78, the branch from young to old mostly because longer transcript or more transcript. The branch from old to young, mostly maybe because some new genome assembly during dating process.		

### 3.3 comparison and explanation between dm6\_ver78 (FB\_2014\_5) and dm3\_ver54

Table 7 detailed comparison between dm6\_ver78 and dm3\_ver54

		dm6_ver78 VS dm3_ver54										
		dm3_ver54										
		0	1	2	3	4	5	6	total	ver78_total	ver78_only	
dm6_ver78	0	10410	22	5	11	17	1	3	10469	12013	1544	
	1	2	143	7	0	2	1	0	155	202	47	
	2	0	11	109	8	3	0	1	132	214	82	
	3	4	0	2	106	9	0	0	121	200	79	
	4	2	1	1	2	217	2	0	225	334	109	
	5	0	0	0	0	3	44	2	49	76	27	
	6	0	0	0	0	0	0	23	23	44	21	
	total	10418	177	124	127	251	48	29	11174	13083	1909	
	ver54_total	11907	220	161	154	283	68	60	12853			
	ver54_only	1489	43	37	27	32	20	31	1679			
young gene identity: (143+109+106+217+44+23) / (11174-10418) = <b>0.87</b>												

Table 8 Different gene detailed information of ver78 and ver54

Explanation of only exist in ver78 or ver54 gene		
	num	reasons
Ver78 only gene	1569	Ver78 has 1569 new coding geneID never exit in ver54(such as FBgn0267727, FBgn0267790) Maybe from ver54 geneID changed
	119	119 gene on common chrom but in ver54 on unuual chrom(such as FBgn0069969, FBgn0069969, FBgn0085638)
	9	These genes have peptide sequence in ver78 but in ver54 have no peptide sequence and non coding gene. such as FBgn0050082, FBgn0083068
Ver54 only gene	1479	1479 coding geneID in ver54discarded or changed geneID in ver78 (such as FBgn0084036, FBgn0069939 changed to FBgn0259677)
	12	12 ste series gene filtered in ver78 but retained in ver54
<b>Note:</b> because I don't know other filter process and corresponding table, I cannot statistic other station number.		
Difference explanation between ver78 and ver54		
version	Branch change	Reasons
Ver54->ver78	1->0	have a longer transcript (FBgn0052106, FBgn0030905, FBgn0033404)
	1->2	1. Transcript number decreased and retain a shorter transcript (FBtr0080240) 2. Maybe due to genome update.( FBgn0036394)
	2->0	Have a much longer transcript (FBgn0085264, FBgn0029724, FBgn0051313, FBgn0085264)
	2->1	have a longer transcript and/or added transcript number (FBgn0003060, FBgn0003060, FBgn0031513)
	2->3	Have a shorter transcript(FBgn0038123) or maybe genome assemble update

	3->0	Have a longer transcript(FBgn0032110, FBgn0034841)
	3->2	Have a longer transcript (FBgn0051949) or maybe new genome assembly (FBgn0004588)
	3->4	Position changed and transcript len more longer(FBgn0043533)
	4->0	Have a longer transcript and more transcript number (FBgn0085358, FBgn0036705)
	4->3	Have a longer transcript (FBgn0051406)
	4->5	Longer transcript(FBgn0052212) or maybe new genome assembly (FBgn0040365)
<p><b>Note:</b> in general, from ver54 to ver78, the branch from young to old mostly because longer transcript or more transcript. The branch from old to young, mostly maybe because some new genome assembly during dating process.</p>		