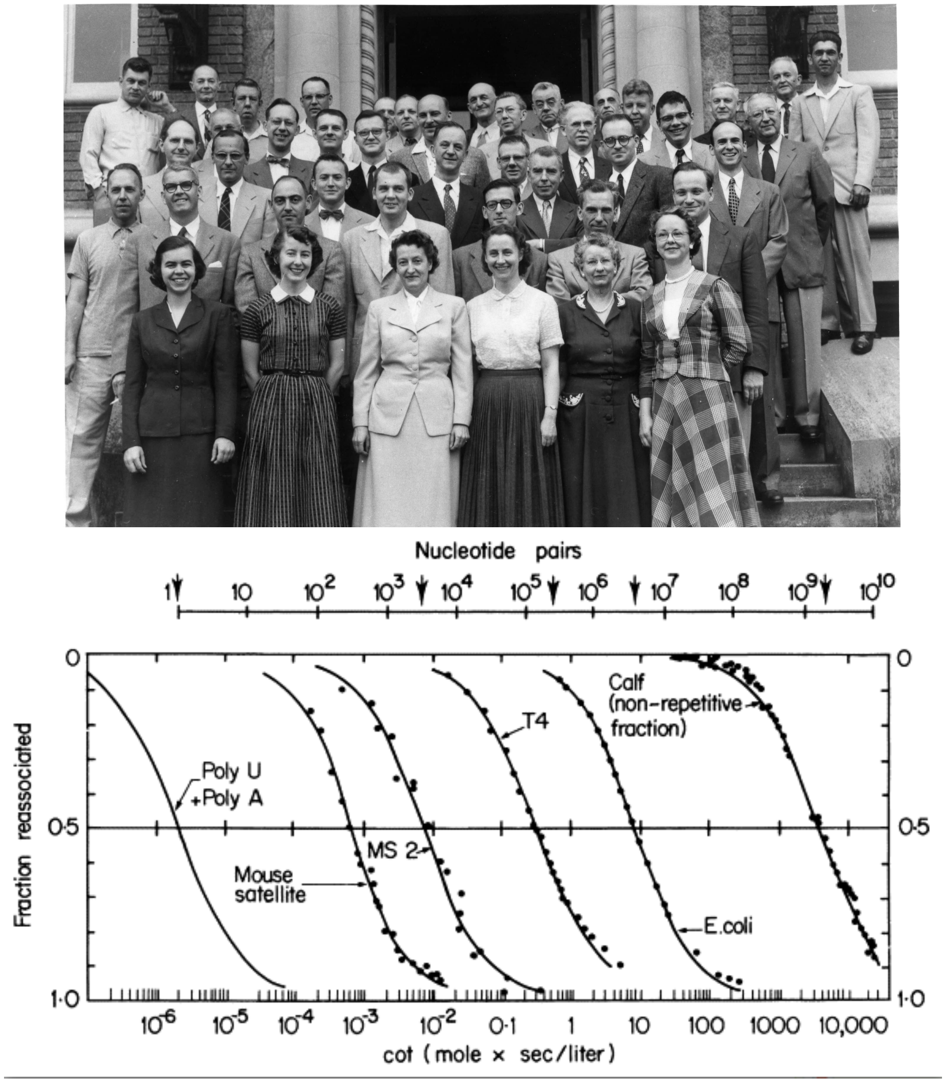
Novel insights into genome structure and evolution as a byproduct of tool generation

modENCODE Symposium
NHGRI Natcher Auditorium
June 21, 2012

Two 20th Century surprises about the genome

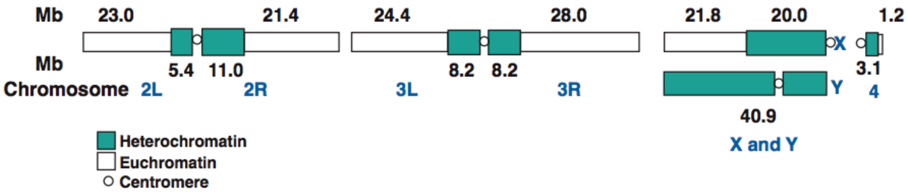
Transposable elements
(1950)



Repetitive DNA
(1960)

Transposons and repeats: the genomic majority

Drosophila



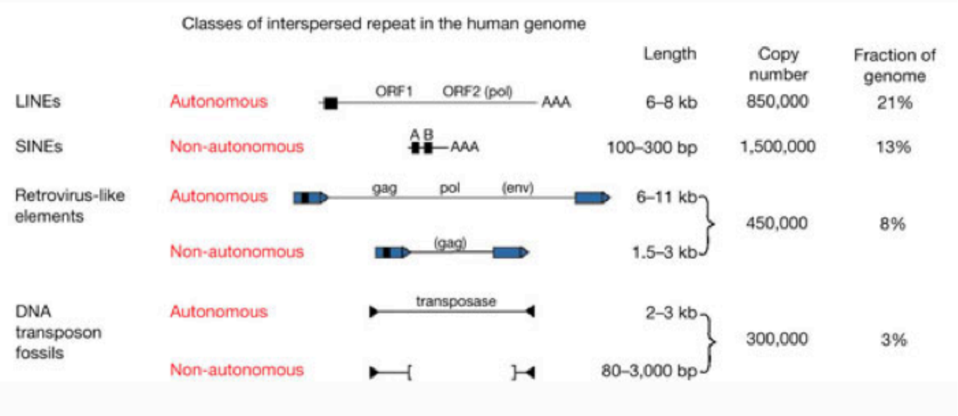
>30% of genome transposon-derived

Full length copies

mariner	0 - 5
piggyBac	0 - 10
P element	0 - 15

The “P element”, a DNA transposon, entered genome recently (~1950), spread throughout world populations

Human



Classes of interspersed repeat in the human genome

			Length	Copy number	Fraction of genome
LINEs	Autonomous	ORF1 ORF2 (pol) AAA	6–8 kb	850,000	21%
SINEs	Non-autonomous	A B AAA	100–300 bp	1,500,000	13%
Retrovirus-like elements	Autonomous	gag pol (env)	6–11 kb	450,000	8%
	Non-autonomous	(gag)	1.5–3 kb		
DNA transposon fossils	Autonomous	transposase	2–3 kb	300,000	3%
	Non-autonomous		80–3,000 bp		

Full length copies

mariner	53,000
piggyBac	500
P element	0*

*12 Thap genes derived from P transposase

Transposons drive human evolution and cancer cell evolution

But we know little about how transposons interact with the genome

Hot and cold spots?

Transposon-specific differences?

Why do transposon-rich regions replicate late in S phase?

Drosophila genome project (1991-2001: NHGRI) and gene disruption project (2001-present: NIGMS)

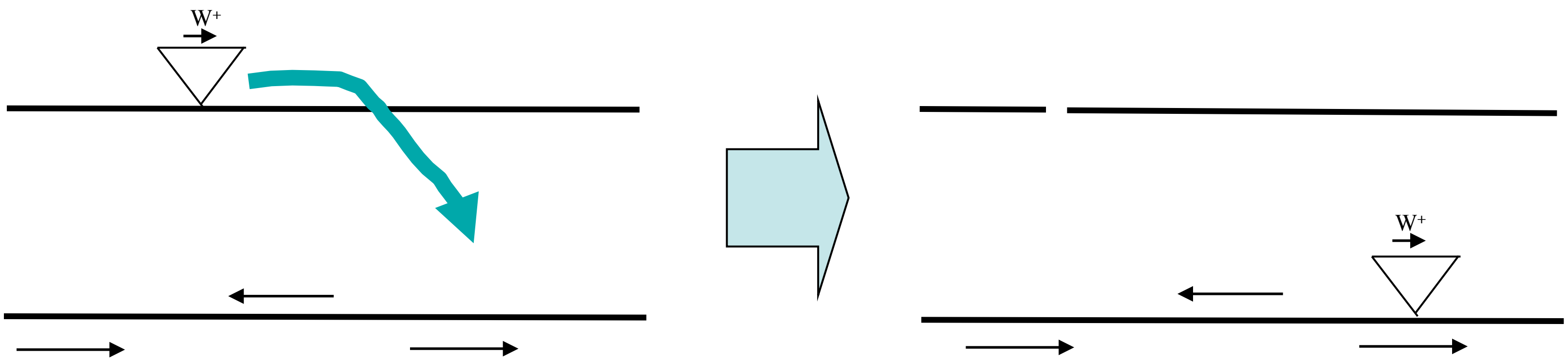
PI's: genome project- Gerry Rubin, Allan Spradling
gene disruption project- Allan Spradling, Hugo Bellen, Roger Hoskins

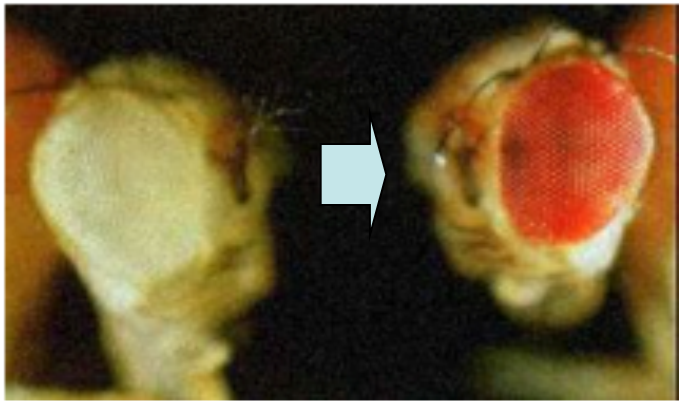
Purpose: generate insertional mutants to determine gene function of all Drosophila genes

Byproduct: the best data on how transposons interact with genomes

A simple experimental paradigm:

Single element jumping screens:





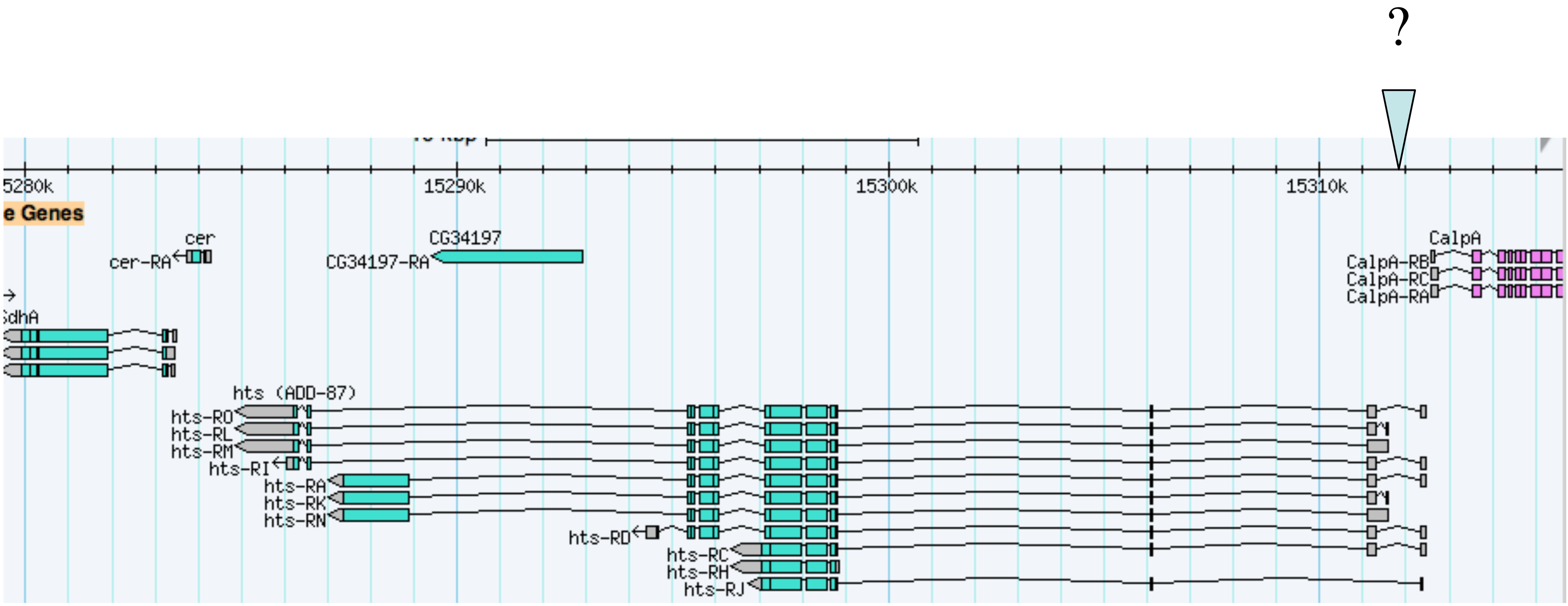
Sequence flank

Advantages of this approach:

Relatively unbiased

Special markers to avoid silencing: yellow, rosy, Su(Var)’s

How do you know which gene(s) are mutated?



Association of insertion lines with genes via their insertion site requires very high quality annotation. Thank you NIGMS and modENCODE for funding annotation.

To understand transposition: must map all insertions from a given starting element

Most screens miss or throw away many insertions; for example, those in suppressive chromatin

Insertions in repetitive DNA cannot always be uniquely mapped

GDP used exceptional care in analyzing insertion sites, and in attempting to identify the correct sites for insertions whose flanks were mostly repetitive

Estimates of insertion in centric heterochromatin probably are the best available, but many isertions were still undoubtedly missed

Mapped events for three transposons

P element	piggyBac	Minos (mariner)
70,593 insertions	17,397 insertions	10,171 insertions

Results for the research community:

>2/3 of all Drosophila genes tagged

Free distribution to the community by BDSC without MTA or strings;

> 250,000 stocks shipped per year from BDSC alone

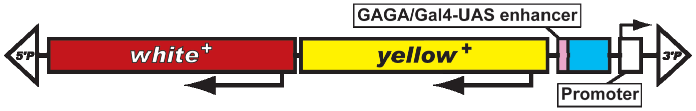
phiC31-based strategy underway for the rest

High quality subset for transposition study

P

18,213 insertions

EY



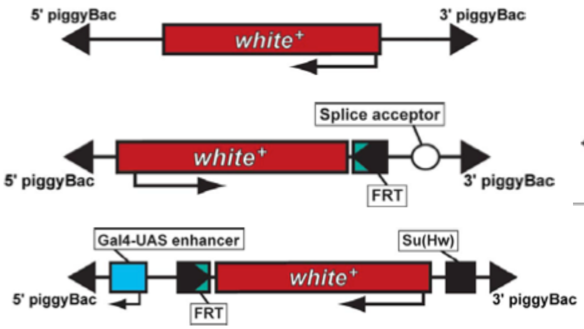
pBac

12,247 insertions

c

e

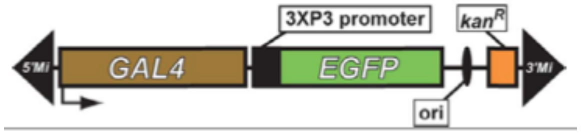
f



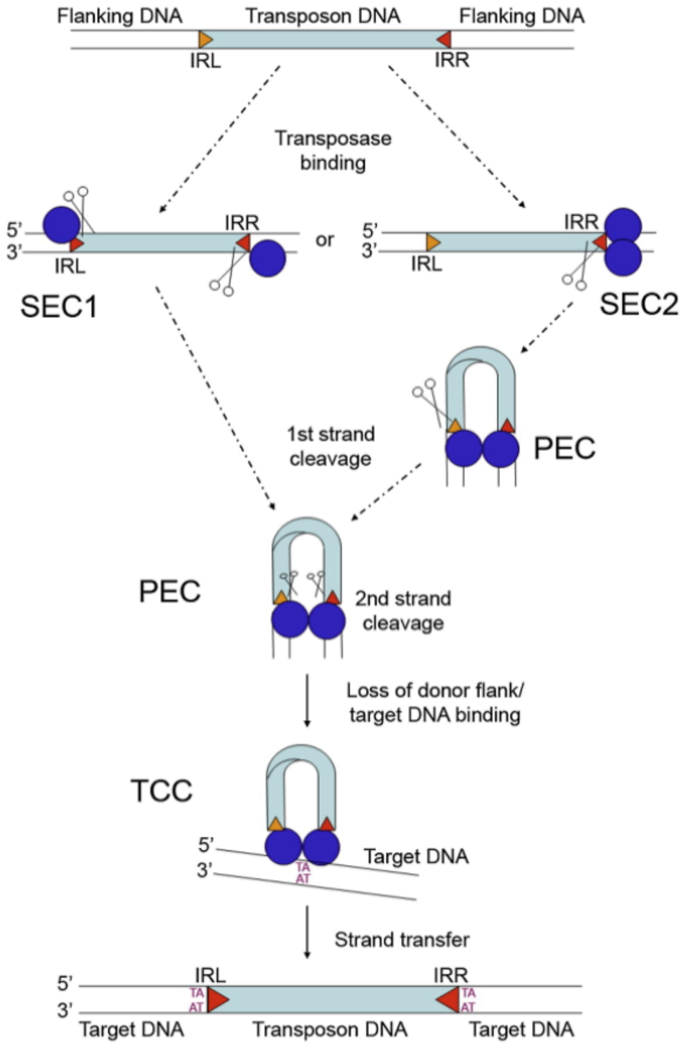
Mar

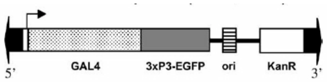
10,171 insertions

MB



Minos: the random transposon



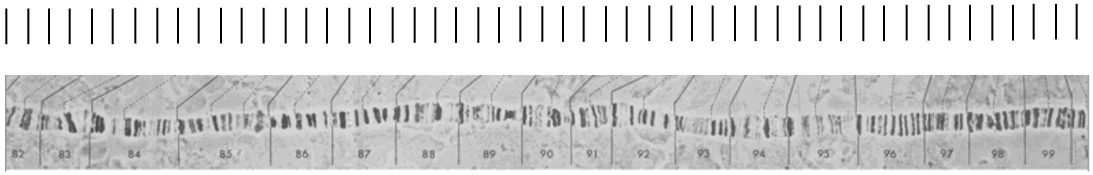
MB 

Minos element
Integrates at TA

Functions efficiently in Ciona, Clostridium, etc.

Human SETMAR gene comprises a SET domain fused to mariner transposase.

Mariner elements transpose more randomly than pBac or P elemnets

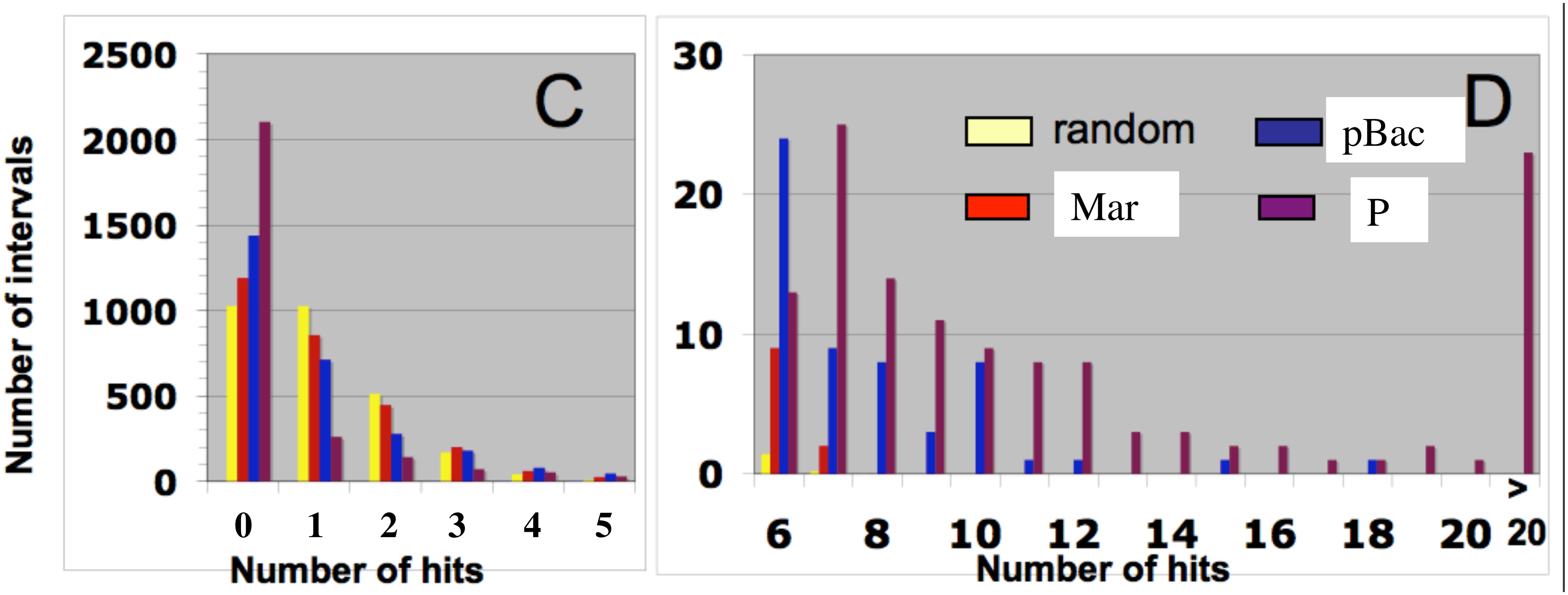


Divide genome into n bins

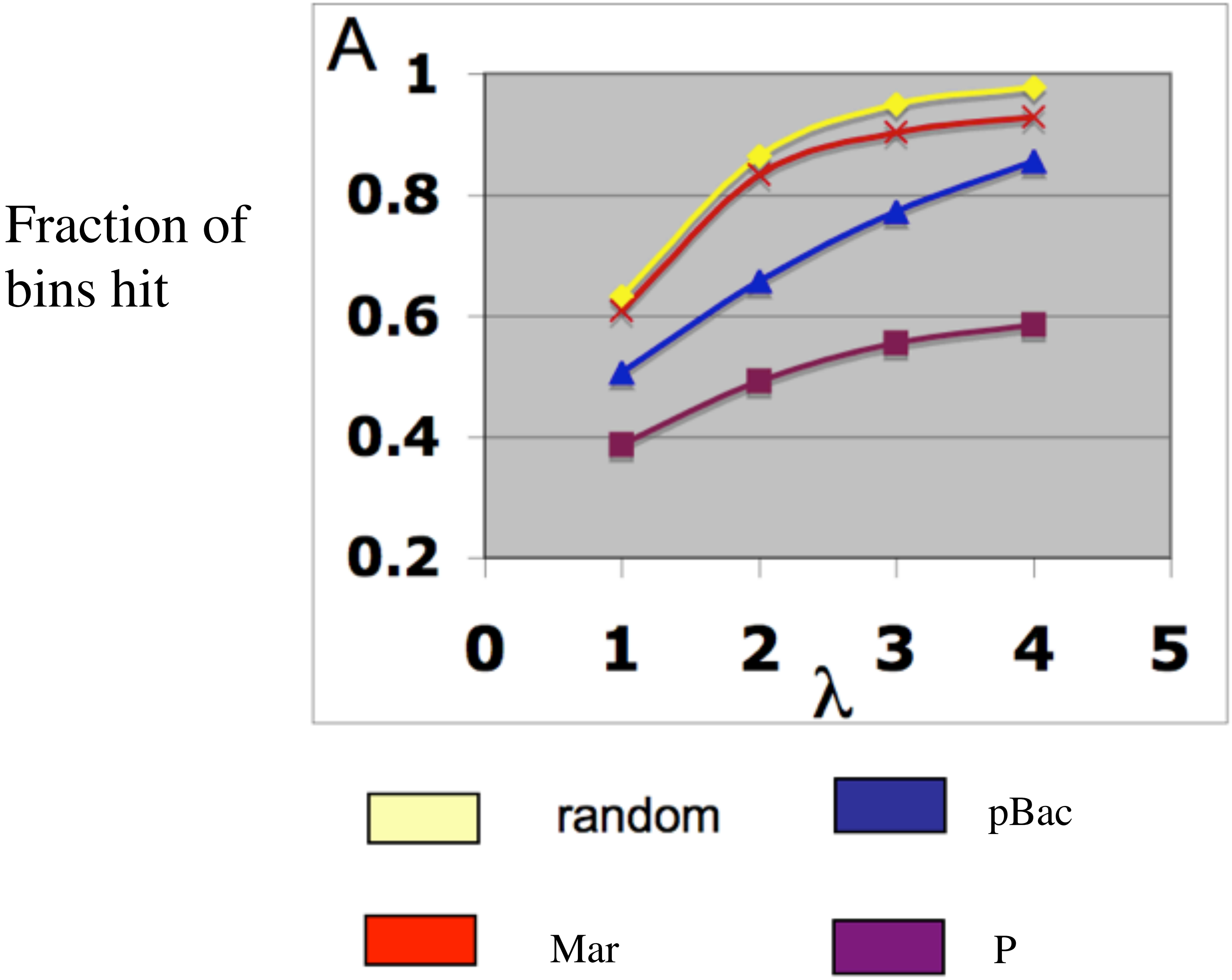
$$P(n,\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$$

$$\lambda = \frac{\text{number of inserts}}{\text{number of bins}}$$

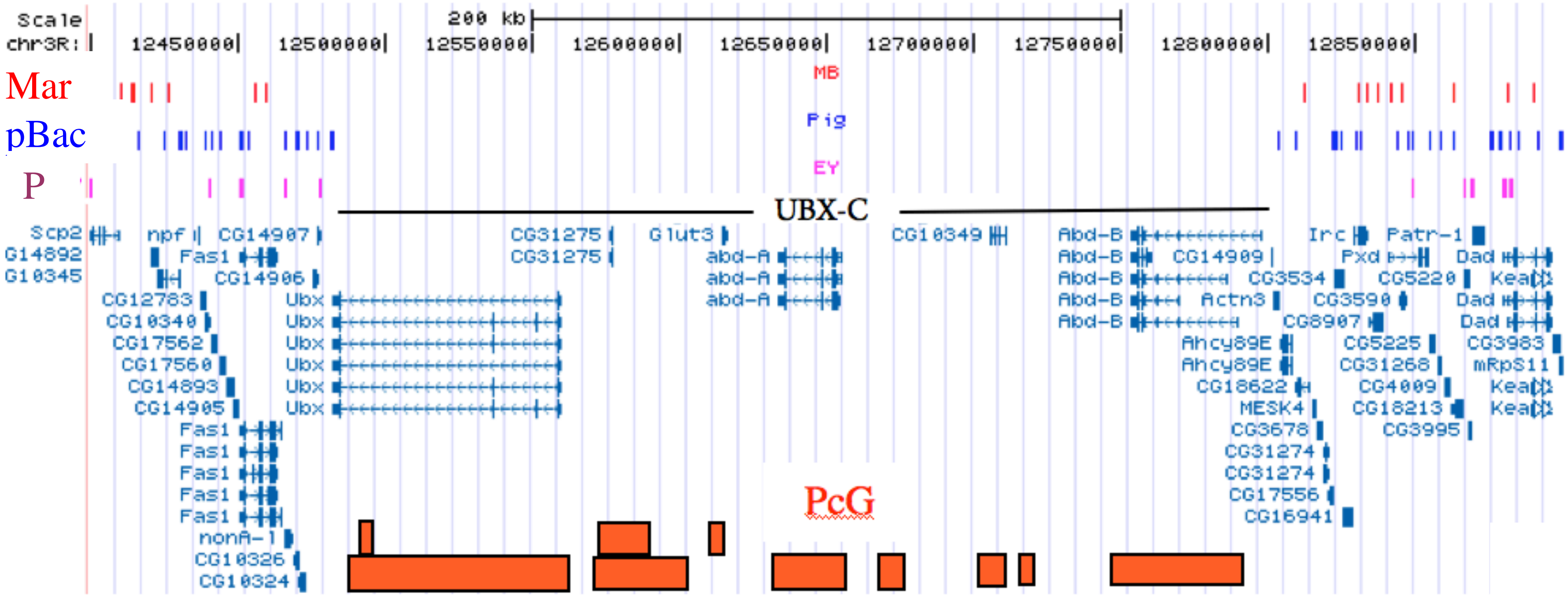
$\lambda = 1$



Major effect on approach to saturation



Deviation from random due to cold spots



Mariner, like all 3 transposons, is recovered less frequently in PcG regulated domains

Many PcG-regulated domains contain few insertions of MB, or the other transposons

Bellen et al. (2011). *Genetics* **188**, 731-43.

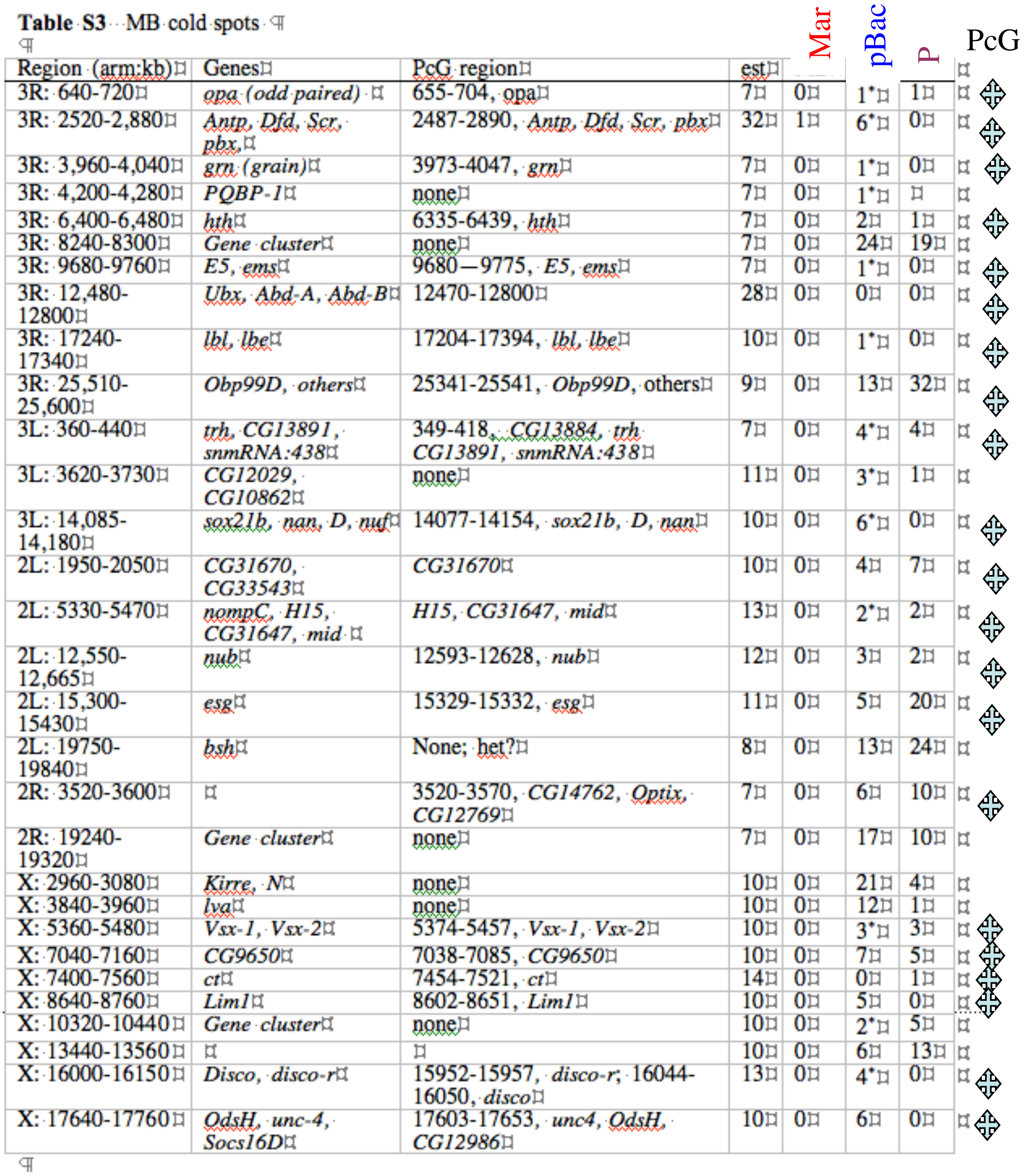
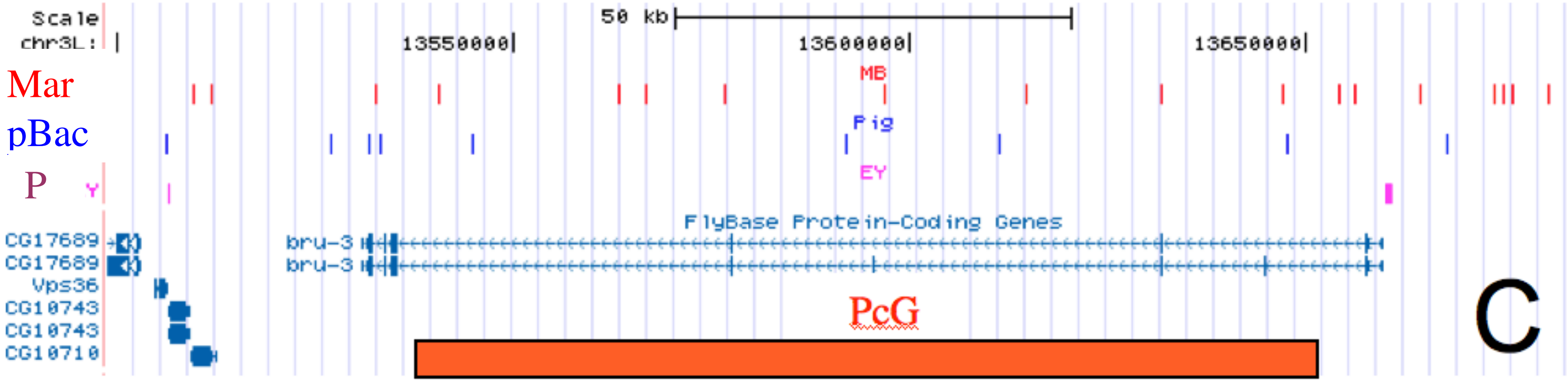


Table S3 MB cold spots

Region (arm:kb)	Genes	PcG region	est	Mar	pBac	P	PcG
3R: 640-720	*opa (odd paired)*	655-704, opa	7	0	1*	1	✥
3R: 2520-2,880	*Antp, Dfd, Scr, pbx,*	2487-2890, *Antp, Dfd, Scr, pbx*	32	1	6*	0	✥
3R: 3,960-4,040	*grn (grain)*	3973-4047, *grn*	7	0	1*	0	✥
3R: 4,200-4,280	*PQBP-1*	none	7	0	1*		
3R: 6,400-6,480	*hth*	6335-6439, *hth*	7	0	2	1	✥
3R: 8240-8300	*Gene cluster*	none	7	0	24	19	
3R: 9680-9760	*E5, ems*	9680—9775, *E5, ems*	7	0	1*	0	✥
3R: 12,480-12800	*Ubx, Abd-A, Abd-B*	12470-12800	28	0	0	0	✥
3R: 17240-17340	*lbl, lbe*	17204-17394, *lbl, lbe*	10	0	1*	0	✥
3R: 25,510-25,600	*Obp99D, others*	25341-25541, *Obp99D*, others	9	0	13	32	✥
3L: 360-440	*trh, CG13891, snmRNA:438*	349-418, *CG13884, trh CG13891, snmRNA:438*	7	0	4*	4	✥
3L: 3620-3730	*CG12029, CG10862*	none	11	0	3*	1	
3L: 14,085-14,180	*sox21b, nan, D, nuf*	14077-14154, *sox21b, D, nan*	10	0	6*	0	✥
2L: 1950-2050	*CG31670, CG33543*	*CG31670*	10	0	4	7	✥
2L: 5330-5470	*nompC, H15, CG31647, mid*	*H15, CG31647, mid*	13	0	2*	2	✥
2L: 12,550-12,665	*nub*	12593-12628, *nub*	12	0	3	2	✥
2L: 15,300-15430	*esg*	15329-15332, *esg*	11	0	5	20	✥
2L: 19750-19840	*bsh*	None; het?	8	0	13	24	
2R: 3520-3600		3520-3570, *CG14762, Optix, CG12769*	7	0	6	10	✥
2R: 19240-19320	*Gene cluster*	none	7	0	17	10	
X: 2960-3080	*Kirre, N*	none	10	0	21	4	
X: 3840-3960	*lva*	none	10	0	12	1	
X: 5360-5480	*Vsx-1, Vsx-2*	5374-5457, *Vsx-1, Vsx-2*	10	0	3*	3	✥
X: 7040-7160	*CG9650*	7038-7085, *CG9650*	10	0	7	5	✥
X: 7400-7560	*ct*	7454-7521, *ct*	14	0	0	1	✥
X: 8640-8760	*Lim1*	8602-8651, *Lim1*	10	0	5	0	✥
X: 10320-10440	*Gene cluster*	none	10	0	2*	5	
X: 13440-13560			10	0	6	13	
X: 16000-16150	*Disco, disco-r*	15952-15957, *disco-r*; 16044-16050, *disco*	13	0	4*	0	✥
X: 17640-17760	*OdsH, unc-4, Socs16D*	17603-17653, *unc4, OdsH, CG12986*	10	0	6	0	✥

A few PcG domains appear exceptional



Hit at expected levels by MB and piggyBac

Conclude: repressive chromatin blocks transposition, and many PcG domains (as assayed in tissue culture, embryo or larval chromatin) are also repressive domains in germ line

Relation to "transposon-free regions" (TFRs) in mammalian genomes

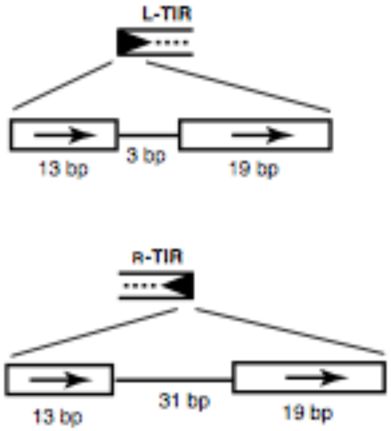
Human TFR	Drosophila ortholog	Dros PcG?	Transposition coldspot?
HOXA4-11	ANT-C, BX-D	Y	Y
HOXB4-6	ANT-C	Y	Y
HOXD8-13	BX-C	Y	Y
DLX5	Distalless	Y	Y
PAX6	ey, so	Y	w
NR2F1	sev	N	N

Transposition in mammals may also avoid PcG domains

Problem: must distinguish lack of transposition with marker suppression

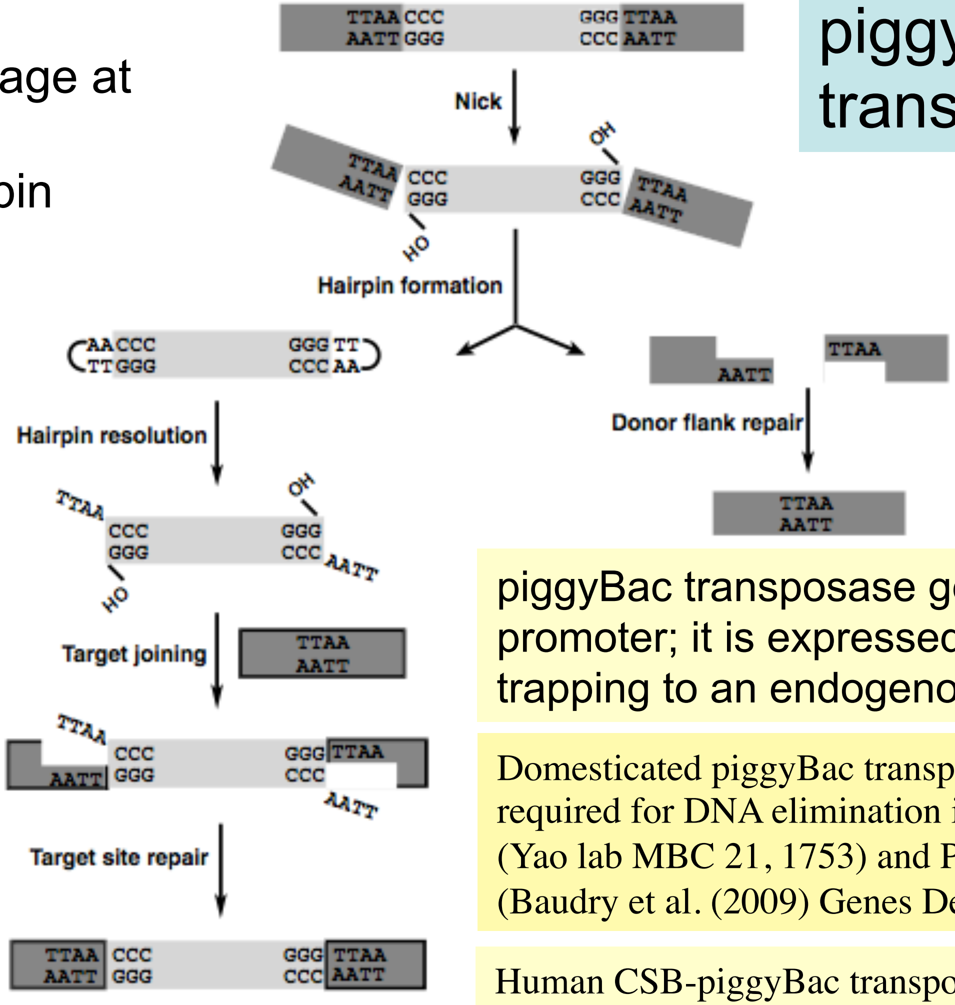
piggyBac

No DNA synthesis required
Transposase catalyzes cleavage at ends and genomic TTAA
Sites, Hairpin formation, hairpin resolution, donor resolution



piggyBac ends

piggyBac transposition

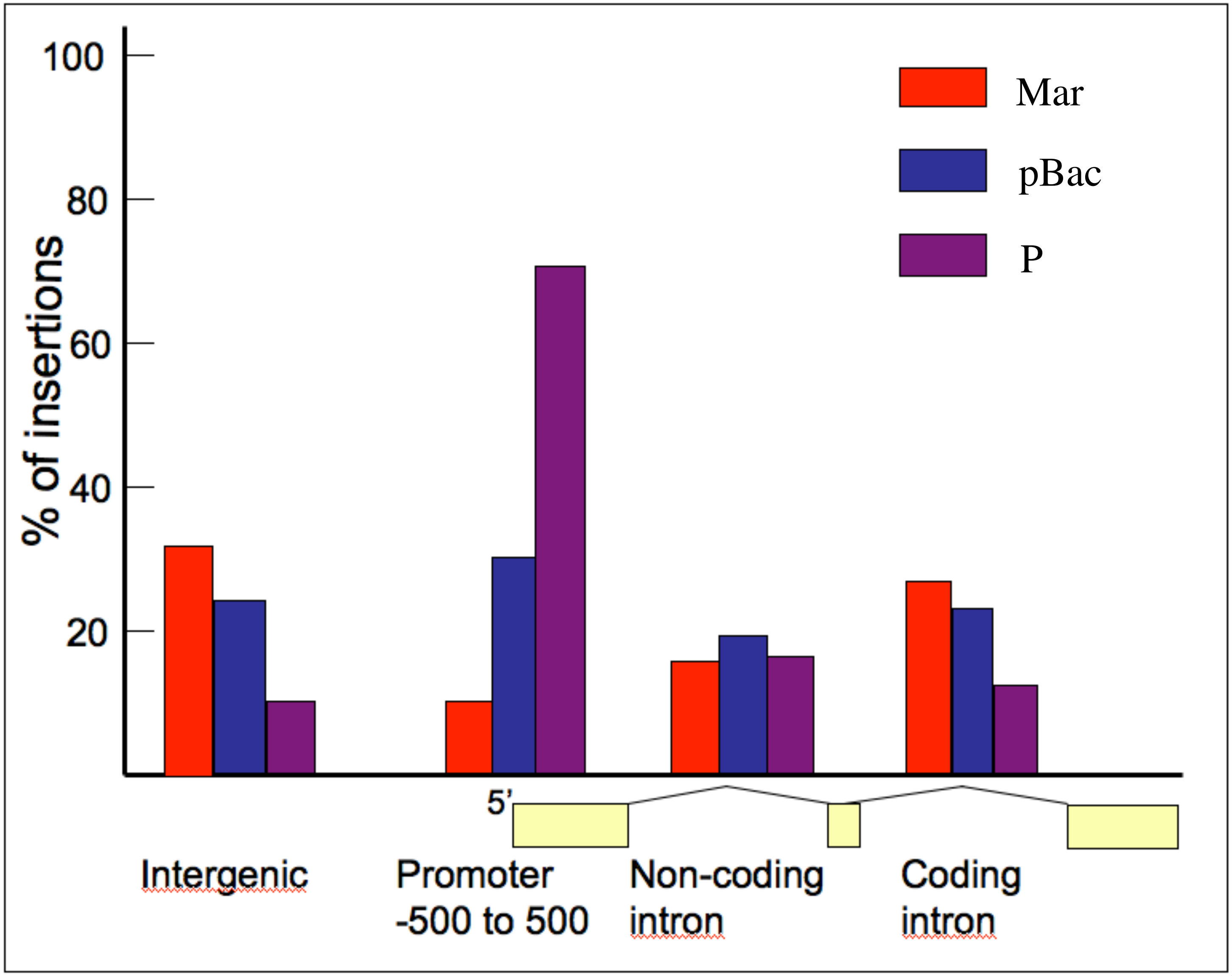


piggyBac transposase gene lacks a promoter; it is expressed via protein trapping to an endogenous gene

Domesticated piggyBac transposase genes are required for DNA elimination in Tetrahymena (Yao lab MBC 21, 1753) and Paramecium (Baudry et al. (2009) Genes Dev. 23, 2478).

Human CSB-piggyBac transposon fusion gene binds 900 defective piggyBac elements in genome. PiggyBac5: transposon encoded by exons

pBac (and P elements) prefer genes and 5' ends



piggyBac cold spots are enriched in membrane proteins and receptors

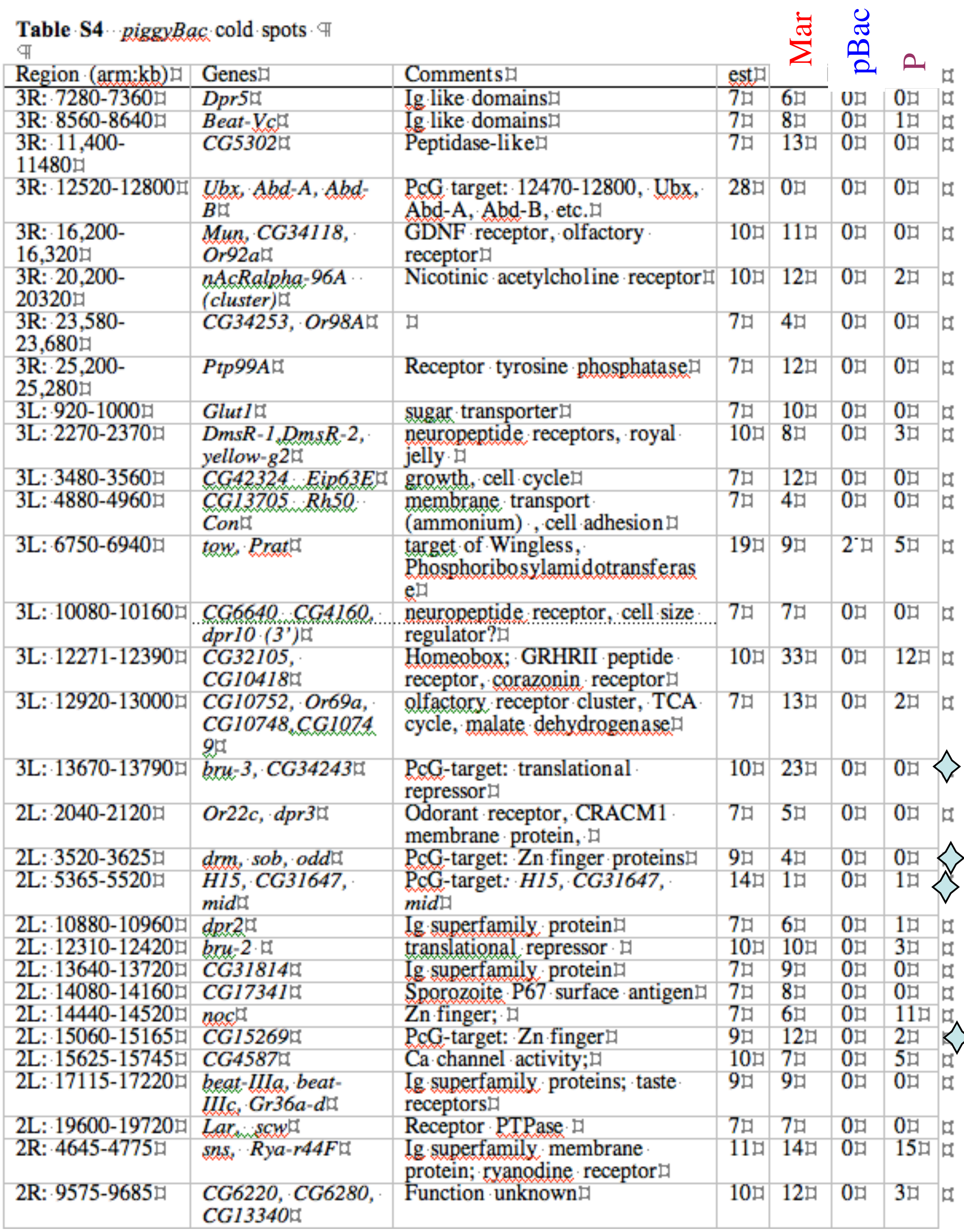


Table S4 *piggyBac* cold spots

Region (arm:kb)	Genes	Comments	est	Mar	pBac	P	
3R: 7280-7360	*Dpr5*	Ig like domains	7	6	0	0	🟡
3R: 8560-8640	*Beat-Vc*	Ig like domains	7	8	0	1	🟡
3R: 11,400-11480	*CG5302*	Peptidase-like	7	13	0	0	
3R: 12520-12800	*Ubx, Abd-A, Abd-B*	PcG target: 12470-12800, Ubx, Abd-A, Abd-B, etc.	28	0	0	0	✧
3R: 16,200-16,320	*Mun, CG34118, Or92a*	GDNF receptor, olfactory receptor	10	11	0	0	🟡🔴
3R: 20,200-20320	*nAcRalpha-96A (cluster)*	Nicotinic acetylcholine receptor	10	12	0	2	🟡🔴
3R: 23,580-23,680	*CG34253, Or98A*		7	4	0	0	🟡🔴
3R: 25,200-25,280	*Ptp99A*	Receptor tyrosine phosphatase	7	12	0	0	🟡🔴
3L: 920-1000	*Glut1*	sugar transporter	7	10	0	0	🟡🔴
3L: 2270-2370	*DmsR-1,DmsR-2, yellow-g2*	neuropeptide receptors, royal jelly	10	8	0	3	🟡🔴
3L: 3480-3560	*CG42324 Eip63E*	growth, cell cycle	7	12	0	0	
3L: 4880-4960	*CG13705 Rh50 Con*	membrane transport (ammonium) , cell adhesion	7	4	0	0	🟡🔴
3L: 6750-6940	*tow, Prat*	target of Wingless, Phosphoribosylamidotransferase	19	9	2	5	
3L: 10080-10160	*CG6640 CG4160, dpr10 (3')*	neuropeptide receptor, cell size regulator?	7	7	0	0	🟡🔴
3L: 12271-12390	*CG32105, CG10418*	Homeobox; GRHRII peptide receptor, corazonin receptor	10	33	0	12	🟡🔴
3L: 12920-13000	*CG10752, Or69a, CG10748,CG10749*	olfactory receptor cluster, TCA cycle, malate dehydrogenase	7	13	0	2	🟡🔴
3L: 13670-13790	*bru-3, CG34243*	PcG-target: translational repressor	10	23	0	0	✧
2L: 2040-2120	*Or22c, dpr3*	Odorant receptor, CRACM1 membrane protein,	7	5	0	0	🟡🔴
2L: 3520-3625	*drm, sob, odd*	PcG-target: Zn finger proteins	9	4	0	0	✧
2L: 5365-5520	*H15, CG31647, mid*	PcG-target: *H15, CG31647, mid*	14	1	0	1	✧
2L: 10880-10960	*dpr2*	Ig superfamily protein	7	6	0	1	🟡
2L: 12310-12420	*bru-2*	translational repressor	10	10	0	3	
2L: 13640-13720	*CG31814*	Ig superfamily protein	7	9	0	0	🟡
2L: 14080-14160	*CG17341*	Sporozoite P67 surface antigen	7	8	0	0	
2L: 14440-14520	*noc*	Zn finger;	7	6	0	11	
2L: 15060-15165	*CG15269*	PcG-target: Zn finger	9	12	0	2	✧
2L: 15625-15745	*CG4587*	Ca channel activity;	10	7	0	5	🟡🔴
2L: 17115-17220	*beat-IIIa, beat-IIIc, Gr36a-d*	Ig superfamily proteins; taste receptors	9	9	0	0	🟡🔴
2L: 19600-19720	*Lar, scw*	Receptor PTPase	7	7	0	0	🟡🔴
2R: 4645-4775	*sns, Rya-r44F*	Ig superfamily membrane protein; ryanodine receptor	11	14	0	15	🟡🔴
2R: 9575-9685	*CG6220, CG6280, CG13340*	Function unknown	10	12	0	3	

🟡 = membrane protein

🔴 = receptor/channel

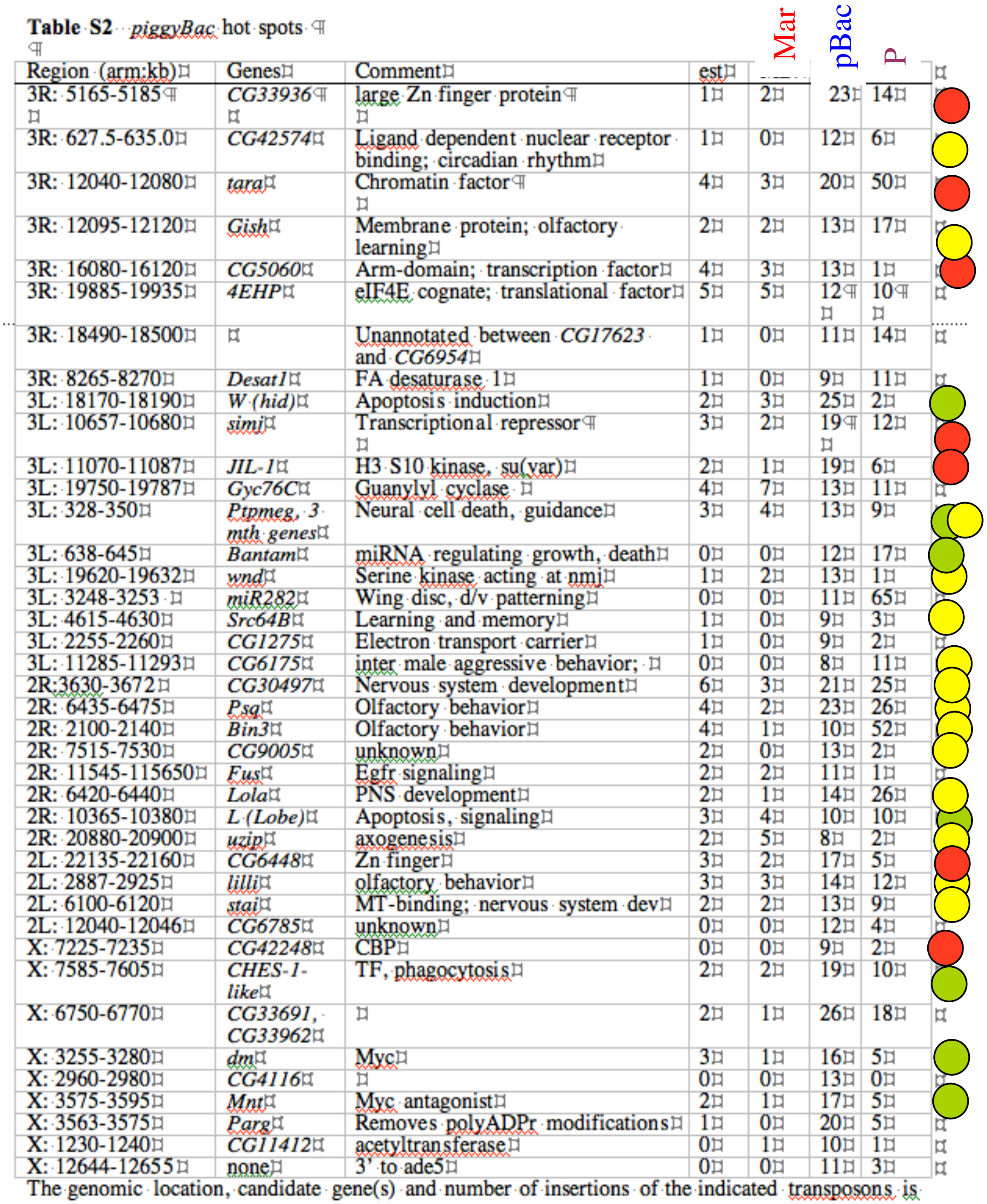


Table S2 *piggyBac* hot spots

Region (arm:kb)	Genes	Comment	est	Mar	pBac	P	
3R: 5165-5185	*CG33936*	large Zn finger protein	1	2	23	14	🔴
3R: 627.5-635.0	*CG42574*	Ligand dependent nuclear receptor binding; circadian rhythm	1	0	12	6	🟡
3R: 12040-12080	*tara*	Chromatin factor	4	3	20	50	🔴
3R: 12095-12120	*Gish*	Membrane protein; olfactory learning	2	2	13	17	🟡
3R: 16080-16120	*CG5060*	Arm-domain; transcription factor	4	3	13	1	🔴
3R: 19885-19935	*4EHP*	eIF4E cognate; translational factor	5	5	12	10	
3R: 18490-18500		Unannotated between *CG17623* and *CG6954*	1	0	11	14	
3R: 8265-8270	*Desat1*	FA desaturase 1	1	0	9	11	
3L: 18170-18190	*W (hid)*	Apoptosis induction	2	3	25	2	🟢
3L: 10657-10680	*simj*	Transcriptional repressor	3	2	19	12	🔴
3L: 11070-11087	*JIL-1*	H3 S10 kinase, su(var)	2	1	19	6	🔴
3L: 19750-19787	*Gyc76C*	Guanylyl cyclase	4	7	13	11	
3L: 328-350	*Ptpmeg, 3 mth genes*	Neural cell death, guidance	3	4	13	9	🟢🟡
3L: 638-645	*Bantam*	miRNA regulating growth, death	0	0	12	17	🟢
3L: 19620-19632	*wnd*	Serine kinase acting at nmj	1	2	13	1	🟡
3L: 3248-3253	*miR282*	Wing disc, d/v patterning	0	0	11	65	
3L: 4615-4630	*Src64B*	Learning and memory	1	0	9	3	🟡
3L: 2255-2260	*CG1275*	Electron transport carrier	1	0	9	2	
3L: 11285-11293	*CG6175*	inter male aggressive behavior;	0	0	8	11	🟡
2R:3630-3672	*CG30497*	Nervous system development	6	3	21	25	🟡
2R: 6435-6475	*Psq*	Olfactory behavior	4	2	23	26	🟡
2R: 2100-2140	*Bin3*	Olfactory behavior	4	1	10	52	🟡
2R: 7515-7530	*CG9005*	unknown	2	0	13	2	🟡
2R: 11545-115650	*Fus*	Egfr signaling	2	2	11	1	
2R: 6420-6440	*Lola*	PNS development	2	1	14	26	🟡
2R: 10365-10380	*L (Lobe)*	Apoptosis, signaling	3	4	10	10	🟢
2R: 20880-20900	*uzip*	axogenesis	2	5	8	2	🟡
2L: 22135-22160	*CG6448*	Zn finger	3	2	17	5	🔴
2L: 2887-2925	*lilli*	olfactory behavior	3	3	14	12	🟡
2L: 6100-6120	*stai*	MT-binding; nervous system dev	2	2	13	9	🟡
2L: 12040-12046	*CG6785*	unknown	0	0	12	4	
X: 7225-7235	*CG42248*	CBP	0	0	9	2	🔴
X: 7585-7605	*CHES-1-like*	TF, phagocytosis	2	2	19	10	🟢
X: 6750-6770	*CG33691, CG33962*		2	1	26	18	
X: 3255-3280	*dm*	Myc	3	1	16	5	🟢
X: 2960-2980	*CG4116*		0	0	13	0	
X: 3575-3595	*Mnt*	Myc antagonist	2	1	17	5	🟢
X: 3563-3575	*Parg*	Removes polyADPr modifications	1	0	20	5	
X: 1230-1240	*CG11412*	acetyltransferase	0	1	10	1	
X: 12644-12655	none	3' to ade5	0	0	11	3	

The genomic location, candidate gene(s) and number of insertions of the indicated transposons is

piggyBac hotspots- enriched for genes involved in growth and behavior?

🟡 = neural development/ behavior

🟢 = growth regulation/apoptosis

🔴 = transcription/chromatin

piggyBac- the good transposon?

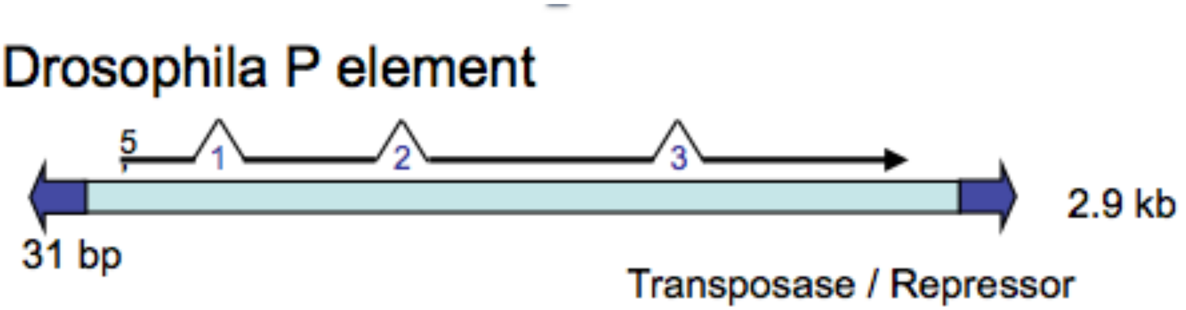
Phylogenetically widespread, hence probably ancient

Domesticated in ciliates to catalyze key events of macronuclear development

Lacks imprecise excision

Has piggyBac adapted its insertional preferences to enhance beneficial and minimize deleterious effects on host?

P element: the selfish transposon?

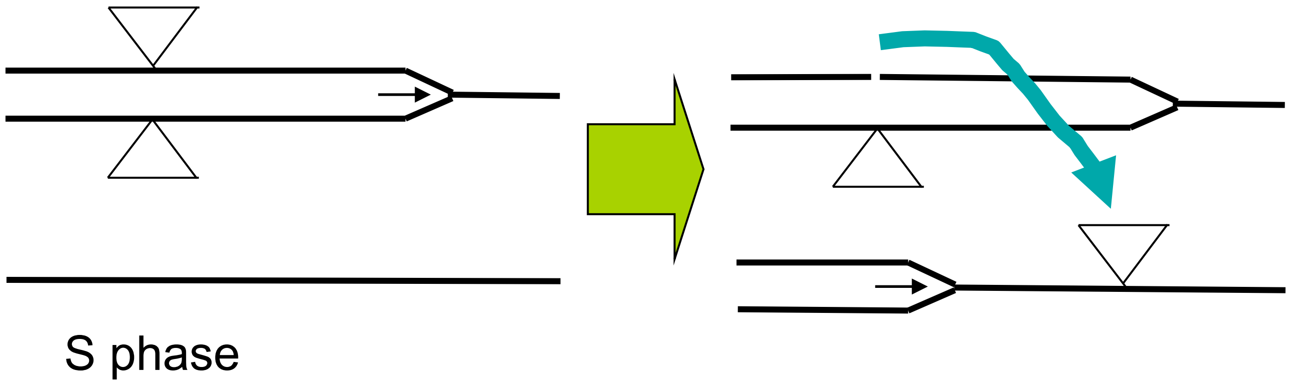


Has rapidly spread throughout D. melanogaster populations worldwide in last 50 years

1 element introduced into a single fly within a laboratory population spreads throughout population in a short time

Conservative DNA transposons require special mechanisms to proliferate

Transposition via cut and paste precludes simple copy number increase

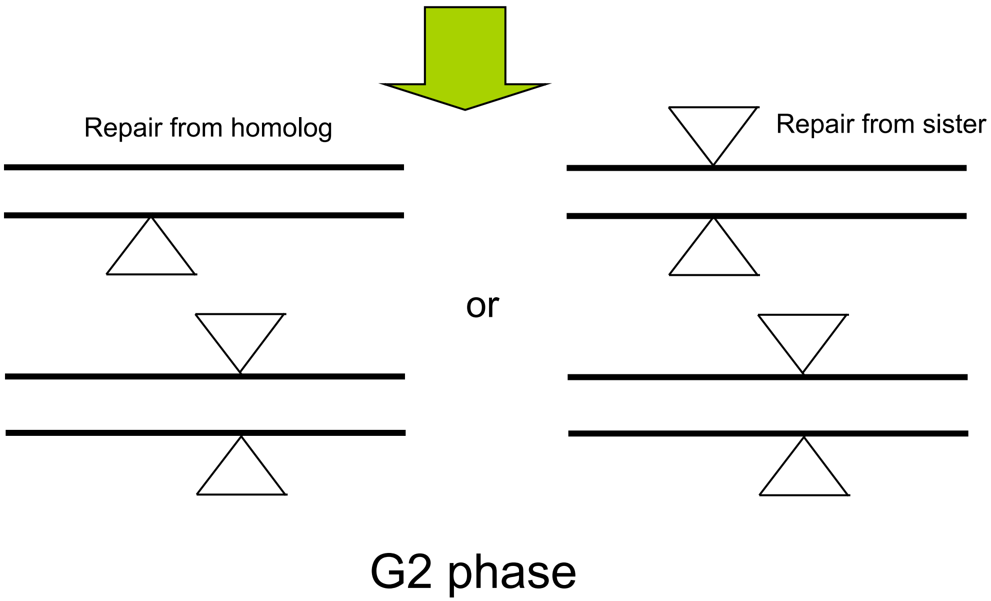


2 potential mechanisms of increase:

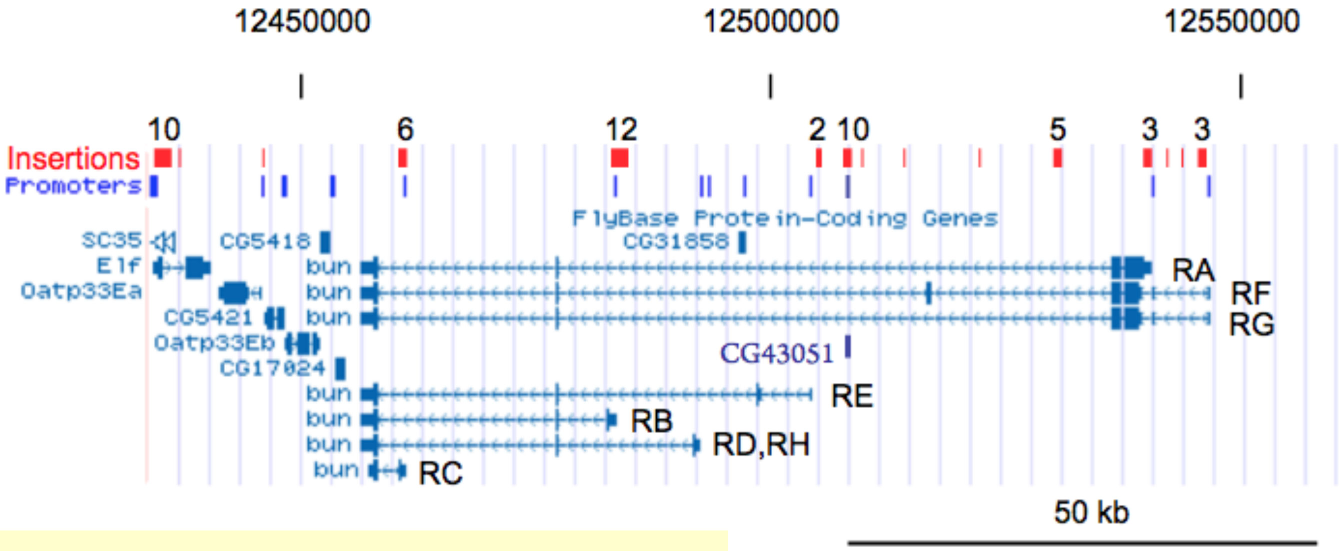
1. Starting site repair (proven)

2. Replication timing (hypothetical)

Limiting transposition to S phase
Limiting transposition to replicated elements
Recognizing unreplicated regions preferentially as target sites

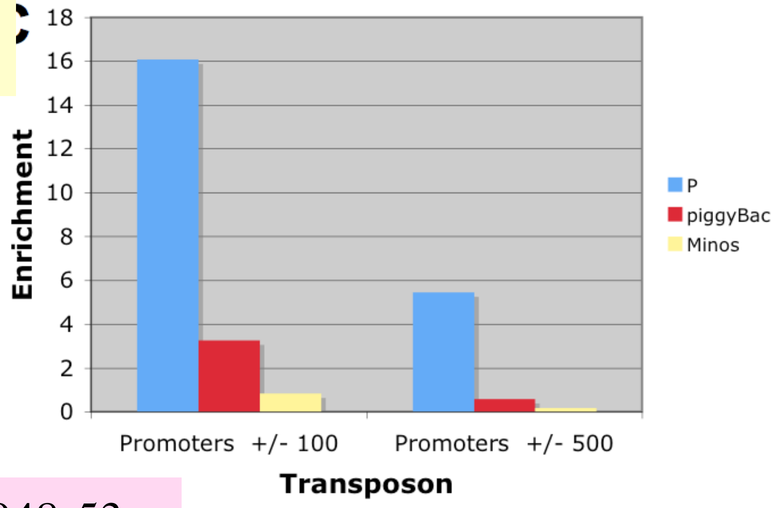


Strong P element promoter preference



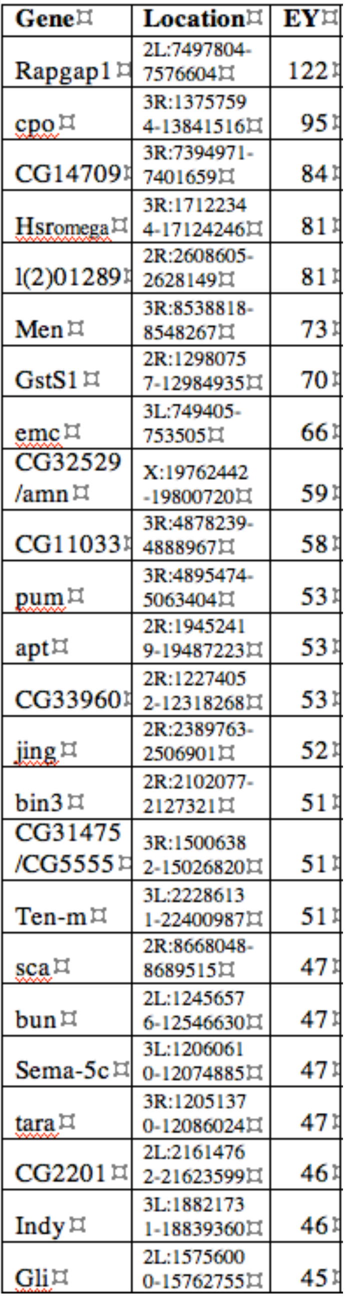
No shared biology between genes that act as hotspots

Almost all tissue-specific clustered genes are coldspots, but so are many other genes



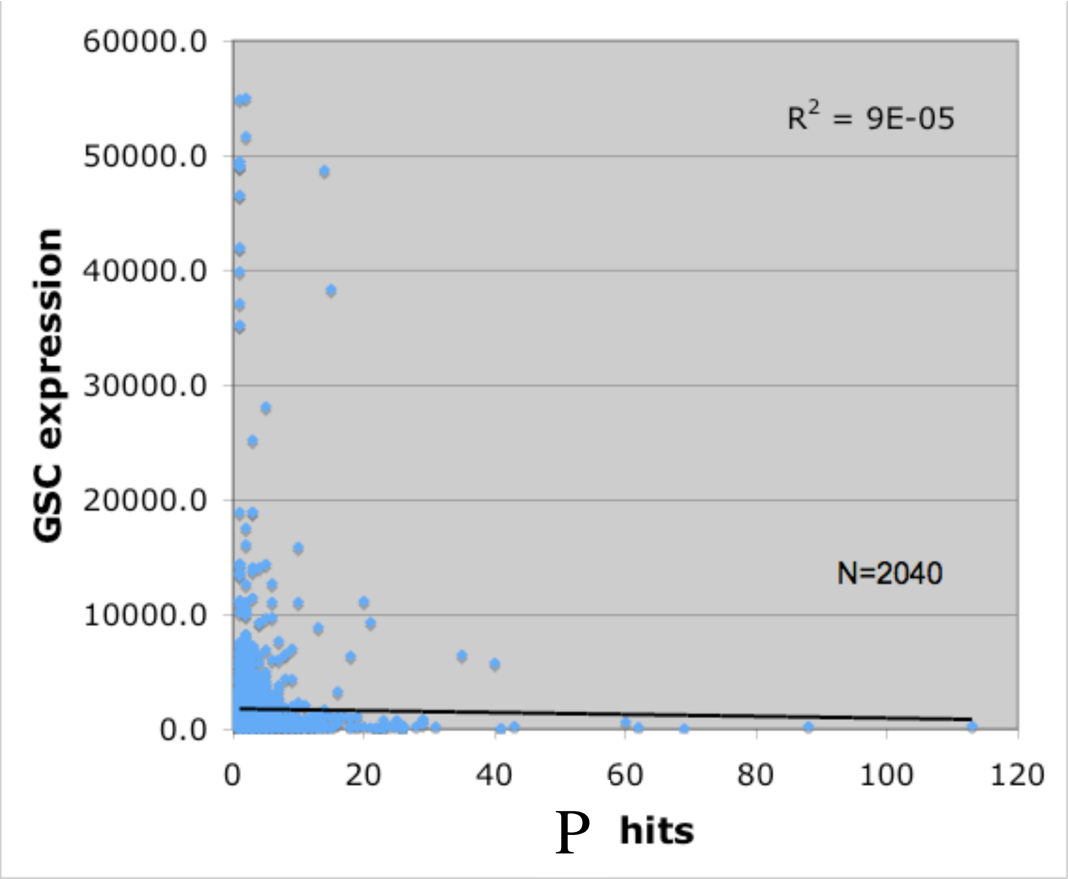
Spradling et al. (2011). *PNAS* **108**, 15948-53.

Top 24 P hotspots

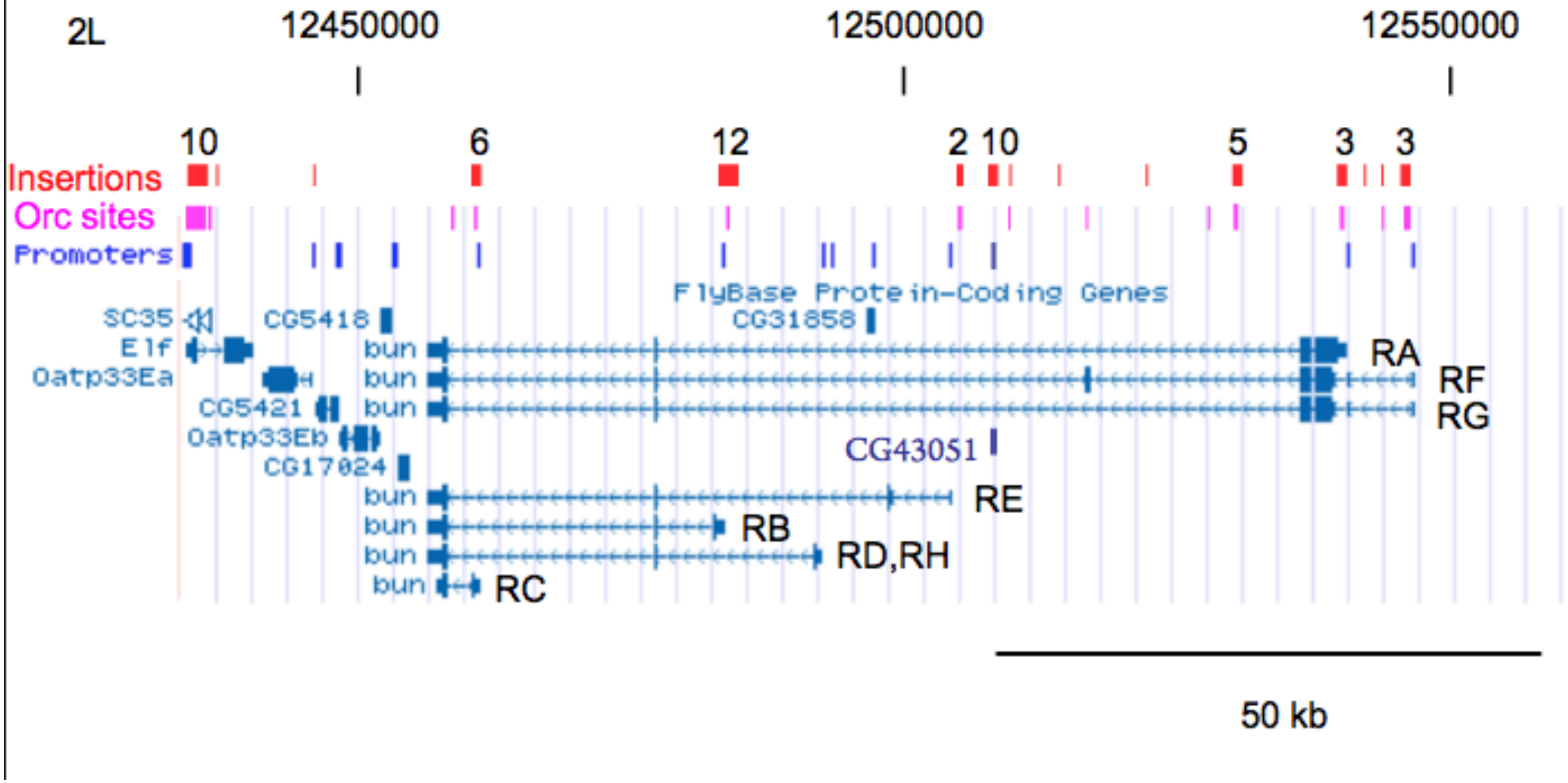


Gene	Location	EY
Rapgap1	2L:7497804-7576604	122
cpo	3R:13757594-13841516	95
CG14709	3R:7394971-7401659	84
Hsromega	3R:17122344-17124246	81
l(2)01289	2R:2608605-2628149	81
Men	3R:8538818-8548267	73
GstS1	2R:12980757-12984935	70
emc	3L:749405-753505	66
CG32529 /amn	X:19762442 -19800720	59
CG11033	3R:4878239-4888967	58
pum	3R:4895474-5063404	53
apt	2R:19452419-19487223	53
CG33960	2R:12274052-12318268	53
jing	2R:2389763-2506901	52
bin3	2R:2102077-2127321	51
CG31475 /CG5555	3R:15006382-15026820	51
Ten-m	3L:22286131-22400987	51
sca	2R:8668048-8689515	47
bun	2L:12456576-12546630	47
Sema-5c	3L:12060610-12074885	47
tara	3R:12051370-12086024	47
CG2201	2L:21614762-21623599	46
Indy	3L:18821731-18839360	46
Gli	2L:15756000-15762755	45

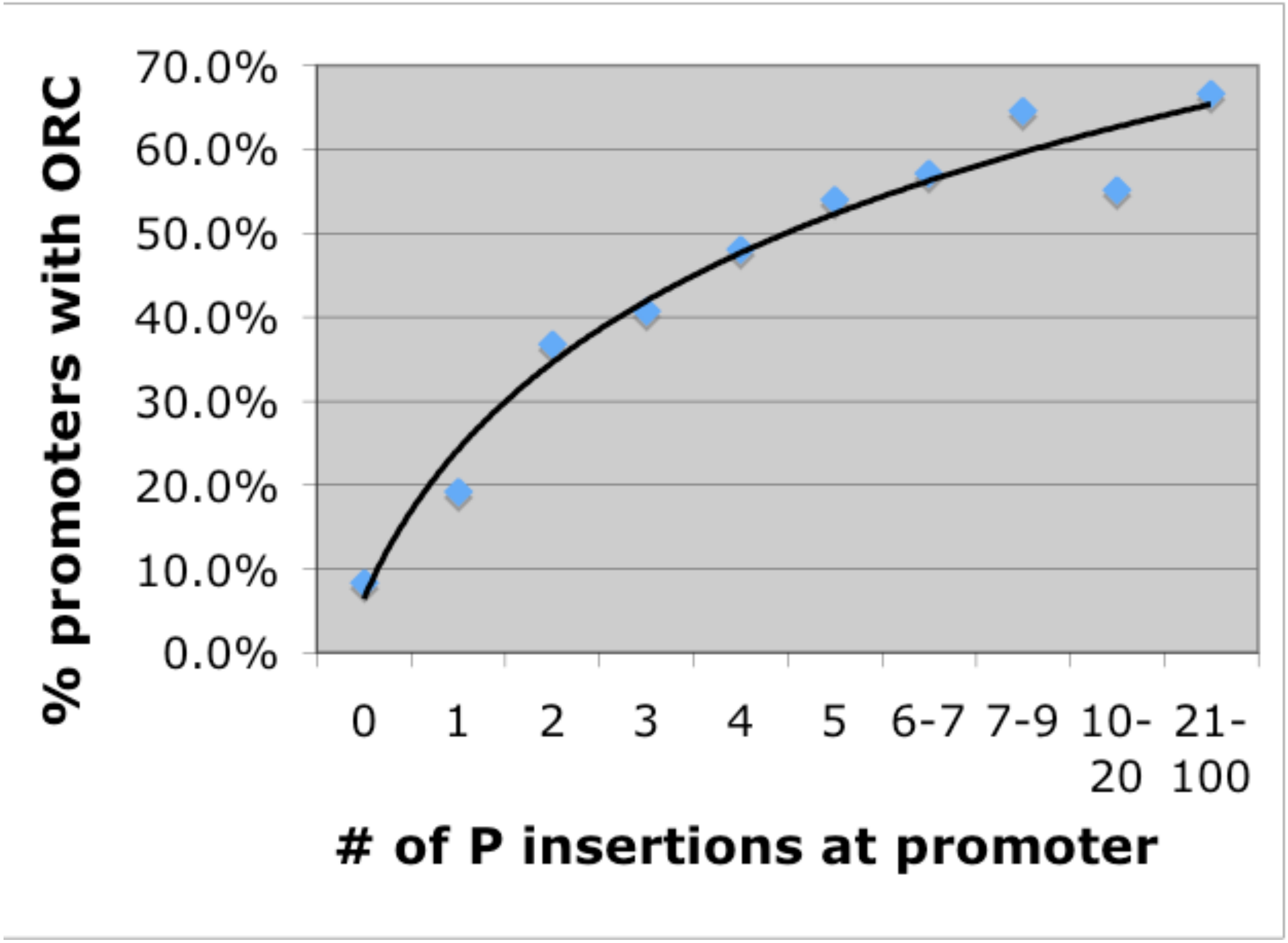
Hotspots are unrelated to transcription in early germ cells



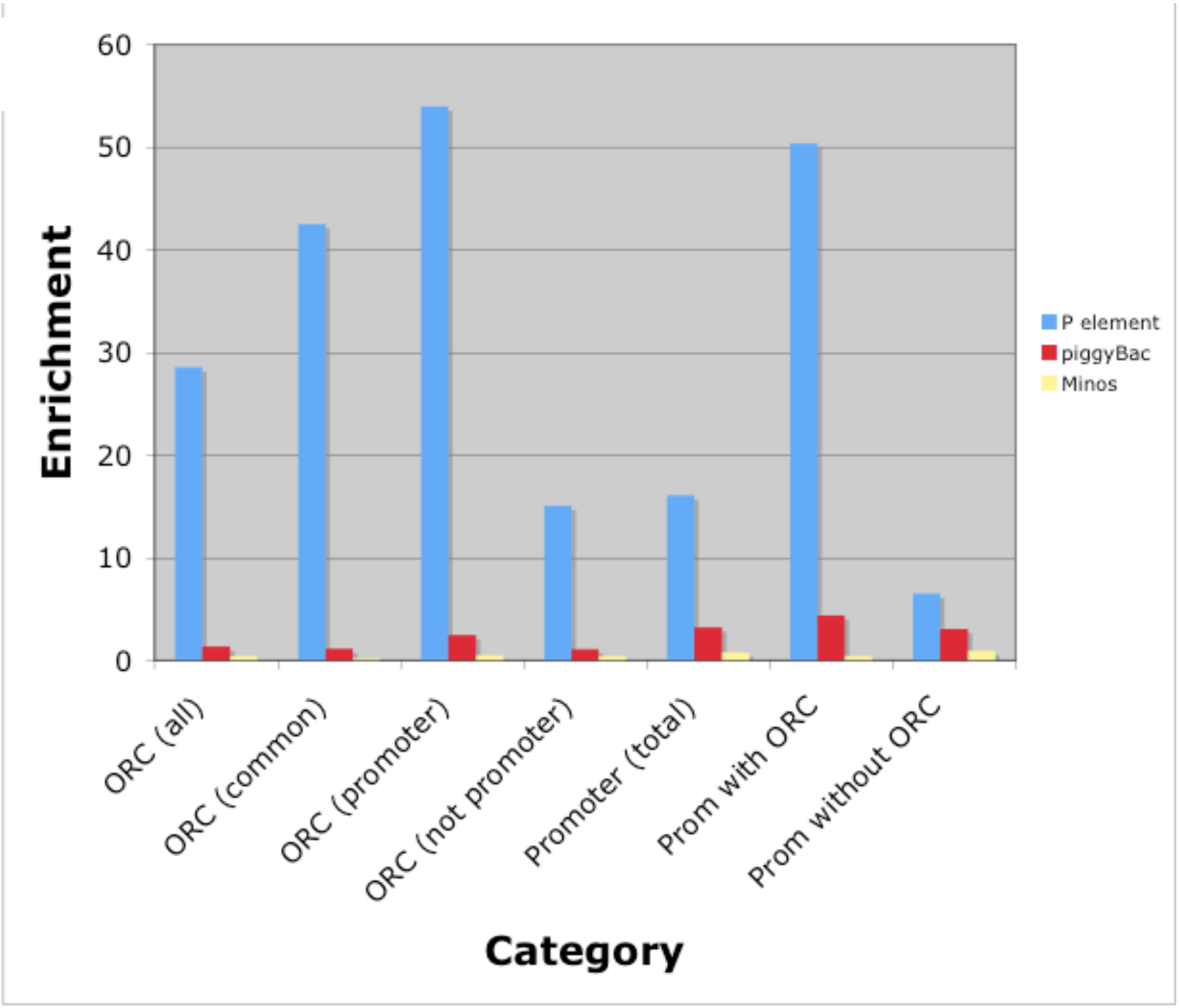
P element hotspots often correspond to replication origins defined by Orc binding



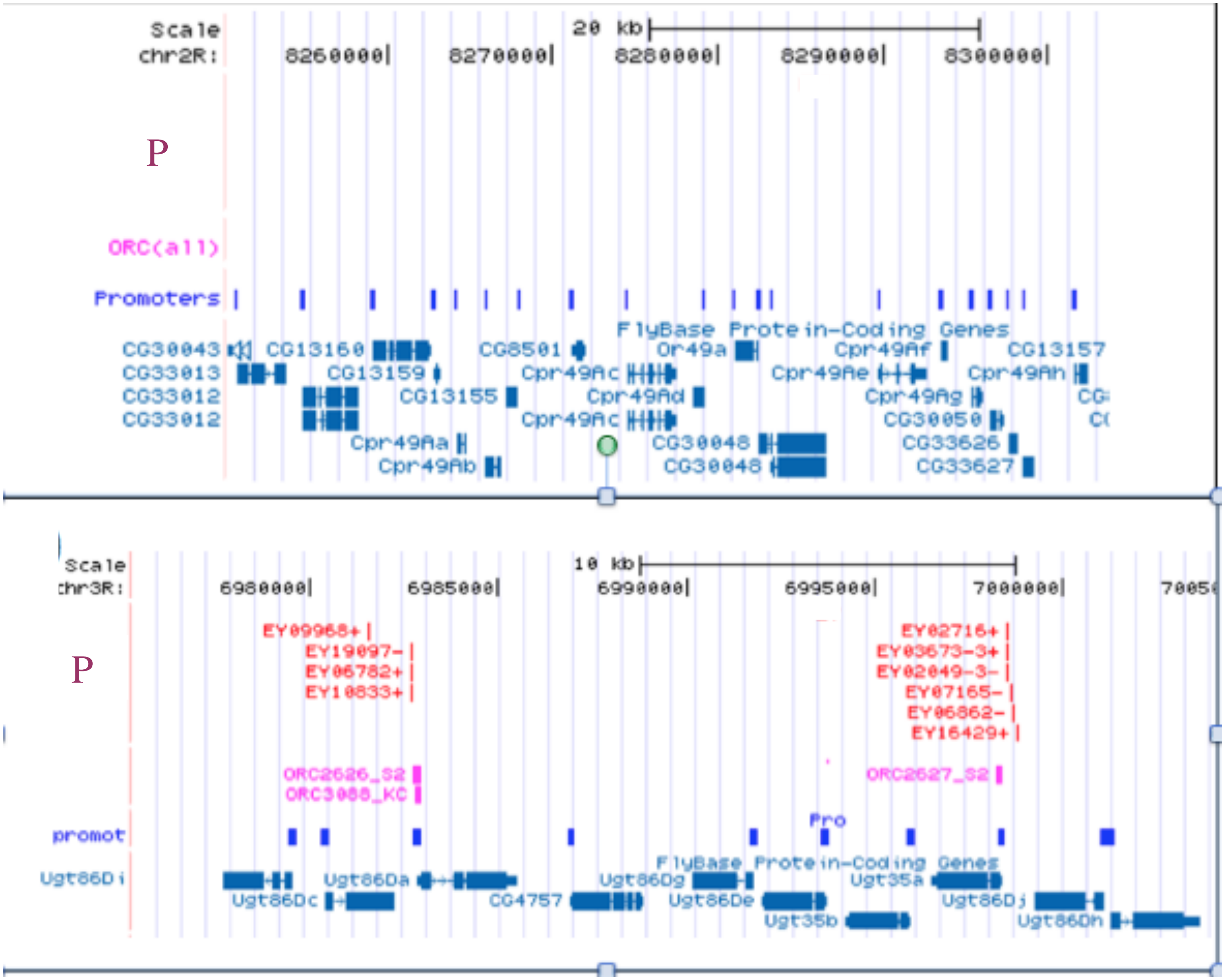
Hotspots are unrelated to transcription in early germ cells



P element enrichment correlates more strongly with origins than with promoters



Tandemly clustered genes usually lack ori's



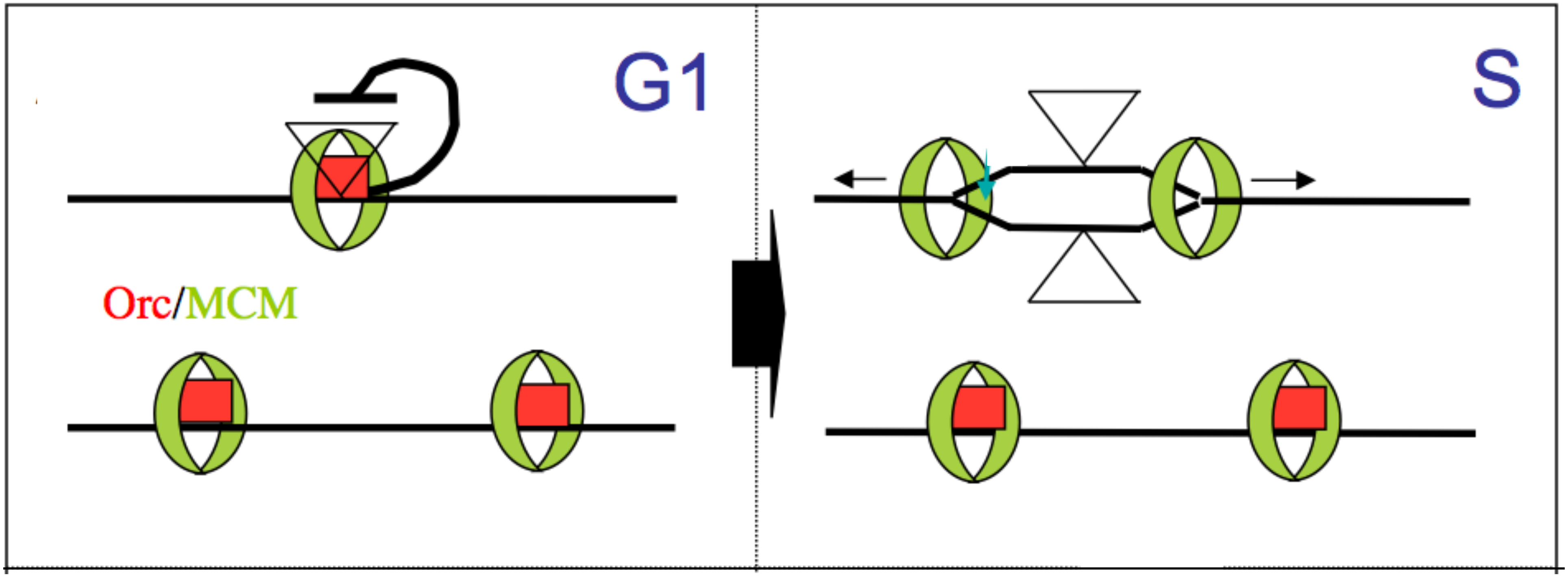
P elements transpose preferentially to replication origins

The origin preference can explain the strong promoter association

The origin preference can explain the lack of transposition in certain classes of genes that lack origins in germ cells

Many origins used in tissue culture cells must also function in early germ cells

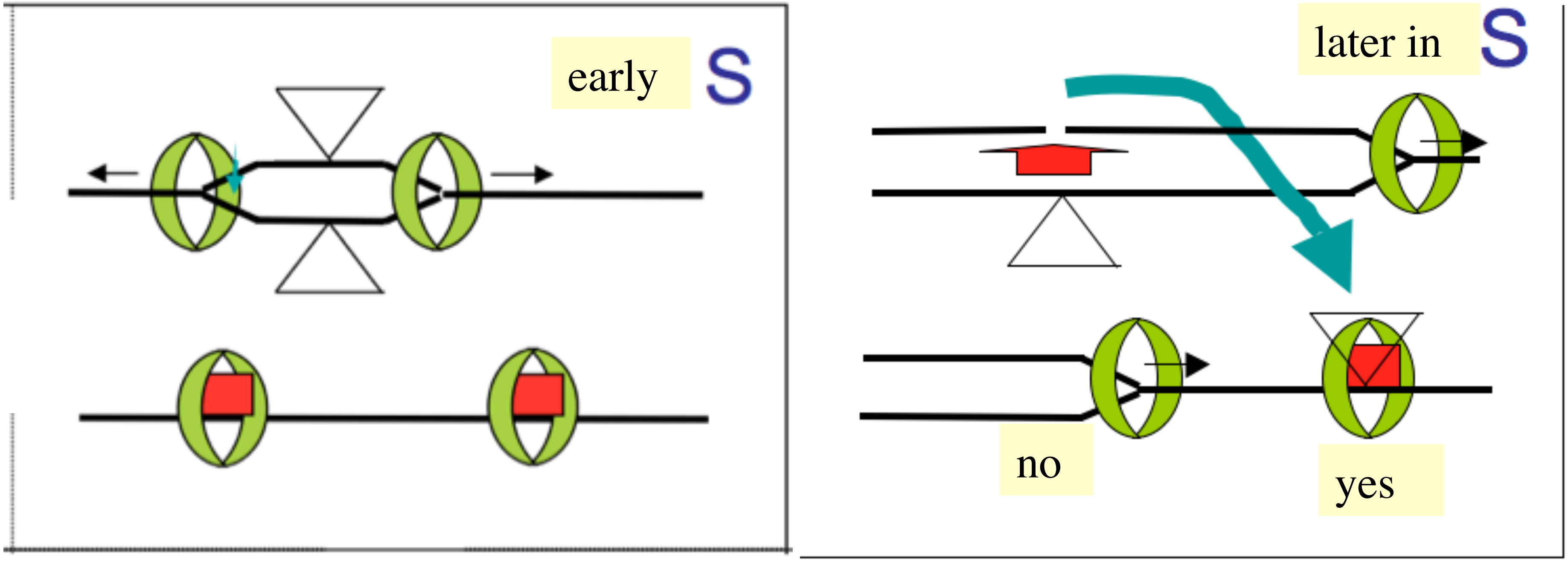
Origin-association may help P elements spread by transposing during S phase



Unactivated origins may repress transposition, limiting movement to replicated regions in S phase

This ensures that a P element-containing homolog will be available for repair

Origin association might also allow P elements to "time" replication



Recognizing part of the pre-initiation complex would distinguish unfired ori's

However, this would require the element to transpose to later firing origins

The selfish drive of transposons to move from early firing to later firing origins may explain why heterochromatin is late replicating

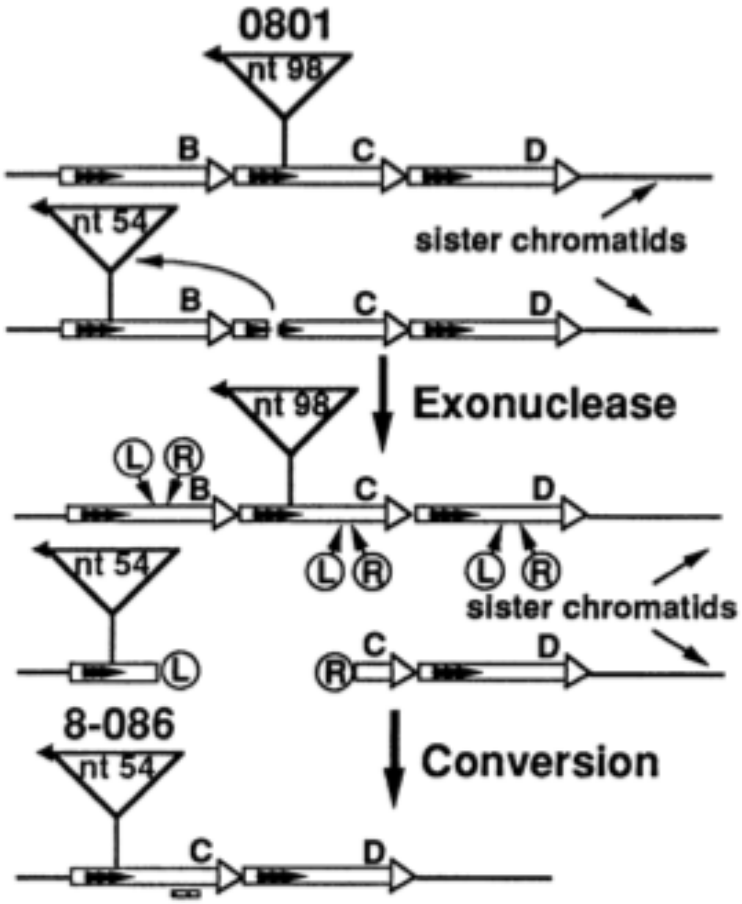
The same benefit would accrue to any transposon, not just to P elements

Transposition into pre-existing elements in these regions could help explain the heterochromatin structure

Genomes might place piRNA loci in late replicating regions to trick new mobile elements into inserting there

High transposon activity could explain the high frequency of tandemly repeated sequences in heterochromatin

Transposon insertion in a tandem repeat stimulates unequal recombination and repeat number changes



Thompson-Stewart et al. (1994) PNAS 91, 9042.

However, an absolute preference for later origins might “trap” active elements

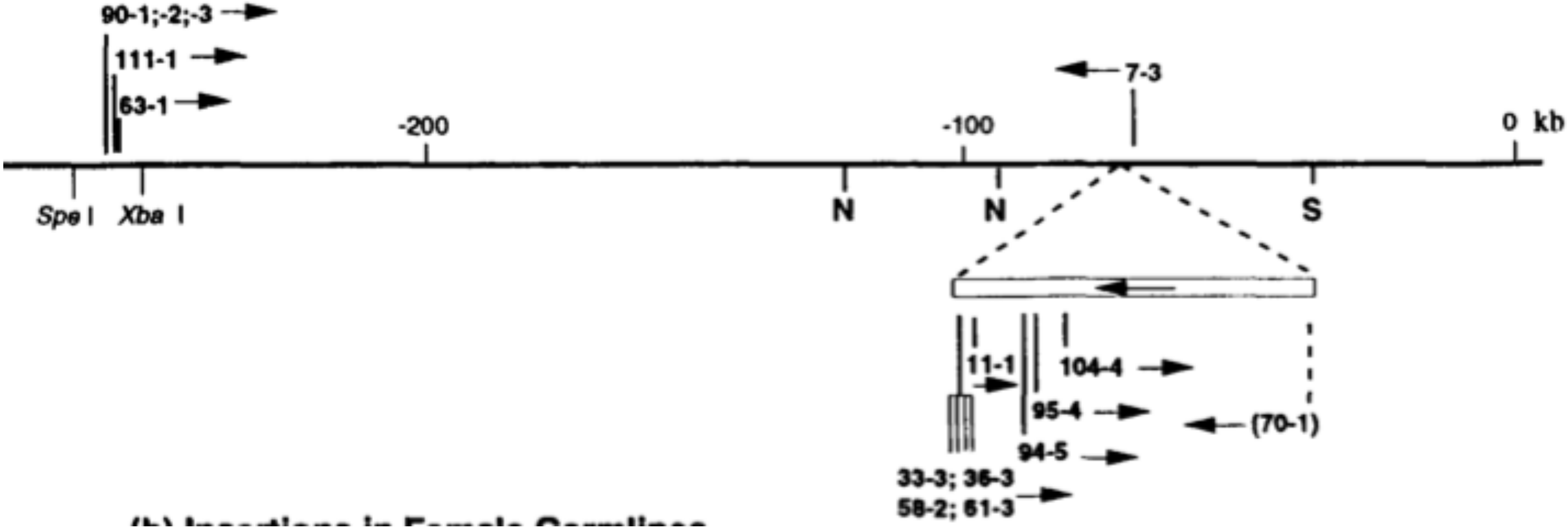
At some frequency, a mechanism is needed to break the cycle, and return elements to earlier replicating regions

Local transposition

Discovered in maize >50 years ago; common to many transposons including P elements

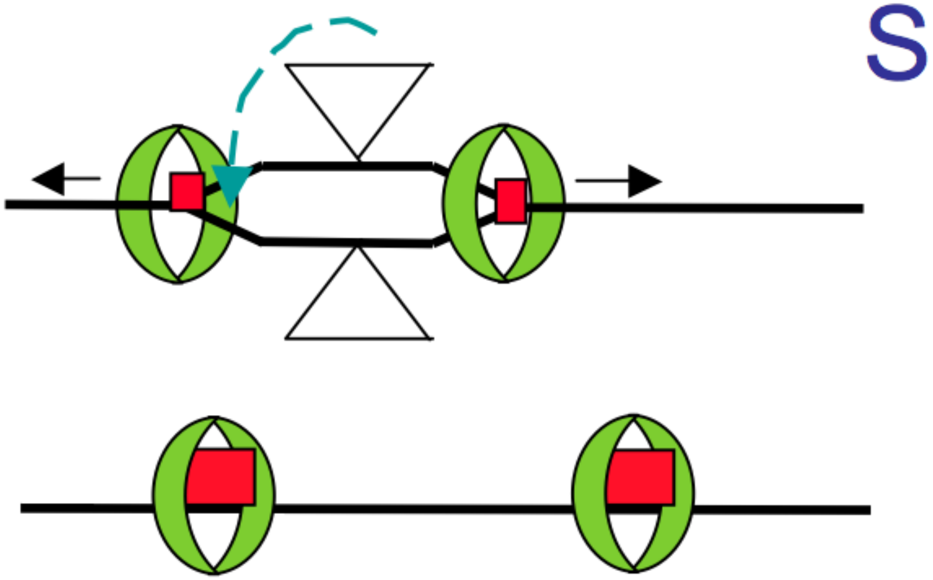
30-70% of transpositions occur near the starting element (0- 200 kb; varies)

Orientation preferences of local jumps



Zhang and Spradling (1993) Genetics 133: 361.

Origin association suggests a simple model of local jumping



For a short time after fork initiation, enough preinitiation proteins may remain at the diverging forks to attract insertion, like an unfired origin

If elements prefer an asymmetric protein, such as PCNA (like Tn7), this would explain the orientation effect

Acknowledgements

Hugo Bellen , Baylor College of Medicine

Roger Hoskins, Lawrence Berkeley Laboratories

Bob Levis, Carnegie Institution

All individuals who contributed to BDGP and GDP during the last 21 years

References: Bellen et al. (2011). *Genetics***188**, 731-43.
Spradling et al. (2011). Proc. Natl. Acad. Sci. **108**, 15948-53.