Fractionation and subfunctionalization following genome duplications: mechanisms that drive gene content and their consequences

[a]Michael Freeling, [b]Michael J. Scanlon, and [c]John F. Fowler

a. Corresponding Author, Department of Plant and Microbial Biology, Univ. California, Berkeley, California 94707

b. Section of Plant Biology, Cornell University, Ithaca, New York 14853

c. Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon 97331

…….

Abstract

A gene's duplication relaxes selection. Loss of duplicate, low-function DNA (fractionation) sometimes follows, mostly by deletion in plants, but mostly via the pseudogene pathway in fish and other clades with smaller population sizes. Subfunctionalization—the founding term of the Xfunctionalization lexicon—while not the general cause of differences in duplicate gene retention, becomes primary as the number of a gene's cis –regulatory sites increases. Balanced gene drive explains retention for the average gene. Both maintenance-of-balance and subfunctionalization *drive* gene content nonrandomly, and currently fall outside of our accepted Theory of Evolution. The "typical" mutation encountered by a gene duplicate is not a neutral loss-of-function; dominant mutations (Muller's lexicon; these are not neutral) abound, and confound Xfunctionalization terms like "neofunctionalization". Confusion of words may cause confusion of thought.

As with many plants, fish tetraploidies provide a higher throughput surrogate-genetic method to infer function from human and other vertebrate ENCODE-like regulatory sites.

Keywords. Whole genome duplication, fractionation, dominant mutation, genome dominance, Gene Balance Hypothesis, subfunctionalization, The Theory of Evolution.

No bullet points. Quote from Conclusion:

"Not only have studies on polyploid fractionation led to reconsiderations of fundamental evolutionary theory, but fractionation in polyploids permits higher-throughput comparative genomic experiments using ENCODE-like data yielding the logical precision expected of genetic analyses."

INTRDODUCTION AND BACKGROUND

We feature, with exceptions, results published from mid-2012 to mid-2015.

*Definitions: distinguishing fractionation and diploidization. "*Fractionation" was coined [1] to differentiate between two different processes that follow whole genome duplications (WGDs): fractionation (mutational loss-of-function of one or the other, but not both, of the newly duplicated genes) and diploidization (adaptations facilitating accurate meiotic segregation). The Salse lab[2] has found, in the grasses, that the phenomenon of genome dominance associated with ancient allopolyploids[3,4] not only influences the sub-genome equivalence of fractionation, but also influences which chromosomes are preferentially rearranged as a part of diploidization. Even so, fractionation is not diploidization. Figure 1 illustrates fractionation of genes, and also fractionation of conserved *cis*-acting elements.

Definitions: Clarifying "subfunctionalization," "neofunctionalization" and similar terms in the Xfunctionalization lexicon. The idea "subfunctionalization" (see Table 1, Row 4 and cartoon inset of Figure 1) was originally called "duplication, degeneration, complementation" by Force and coworkers [5] to explain why too many genes were retained as homeologous pairs after the boney fish tetraploidy. After subfunctionalization, it takes the pair to express the ancestral function at all the right times and places [5-7].

"Subfunctionalization" is one of the customized terms (the "Xfunctionalization" vocabulary [8]) describing the mutational fate of gene duplicates, given that mutants are (generally) neutral because, presumably, the wild-type gene of the pair covers the recessive phenotype [9]. Connant and Wolfe [10] accurately defined "subfunctionalized" and "neofunctionalized" gene homeologs containing a mutant by using examples. However, more generalized treatments were more complete, less accurate, and sometimes confusing from the perspective of a mutant's mechanism of action. For example, Innan and Kondrashov's lexicon [8] describes how either purifying or positive selection might treat each paired gene arrangement of "loss-of-function mutations". Table 1, Row 4, Column 5, compares cartoons of subfunctionalization with nonfunctionalization; "-" is neutral loss-of-function. When mutants are recessive, and neutral or nearly so, the Xfunctionalization terminology is useful.

Mutations precede selection. In 1932, H.J. Muller [11] categorized dominant mutants (Table 1, Column 2) based on how *morphological phenotypes* were affected in a background with alternative dosages of segments carrying the wild-type allele. Dominant mutants are not expected to be neutral . Muller's terminology can also be usefully applied at the level of RNA abundance and distribution: *expression phenotypes*. Deletion of a silencer *cis* site or obstructing looping of an enhancer onto its gene's proximal promoter are two ways to model a hypermorph (over-expression; Table 1, Row 4; Fig. 2A and 2D). Were there an associated hypermorph mutant phenotype, it would be dominant, not neutral. Applying the Xfunctionalization vocabulary to dominant mutants is problematical.

For example, it seems reasonable that a "neofunctionalization" requires a "neomorph" dominant mutant (Table 1, Row 8). Neomorphs (one type of dominant mutant) are something new, and are thus not responsive to any number of wild-type alleles. Eichenlaub and Ettwiller [12] studied the result of the massive (75%) fractionation of duplicates following the teleost fish lineage tetraploidy, and found rare neomorphs produced by a neofunctionalization-type fractionation. In these rare cases, one homeolog was deleted, leaving only a few fragments of exon *in situ*. These retained, conserved sequences had enhancer activity, but it was shown that the original exon DNA did not. Thus, this type of fractionation generated a new enhancer *for a neighboring gene*(s), a variation on the classical Lewis scheme[9] (duplication, repression of one gene which accumulates mutants, derepression, and a chance for a rare neomorph). Recent work [13] following expression of maize homeologs in vegetative leaf blades versus husk sheaths (these being different parts of leaves specialized for photosynthesis and kernel protection, respectively) reported that a remarkable 13% of the pairs were "regulatorily neofunctionalized" and that such pairs were necessarily expressed to different levels. Since dominant hypermorphs (Table 1, Row 4) are expected to be far, far more common than neomorphs, we think that the authors, problematically, used the word "neofunctionalization" in a way that does not demand that anything new has evolved, and called pairs with a dominant over-producer mutants "neofunctionalized". These dominant expression mutations are likely something expected, each like a deletion of an element with a negative function (like a silencer).

Fractionation in fish and plants. Teleost fish, with their lineage-specific tetraploidy, may mutate their *cis*-elements faster than non-polyploid control lineages [14]; we know of no other data on special polyploid mutational mechanisms. The fractionation mechanism is known for maize[15] and *Brassica rapa* [16] genes, and for *Brassica cis*-acting sites and G-box motifs [17]: deletion between short tandem repeats caused by intrachromatid recombination, *not* the pseudogene pathway. While pseudogenes do appear in specific regions of plant genomes [18,19], these broken reading frames were not found at fractionated loci in those plants studied , and plant pseudogenes are not scattered about euchromatin as they are in humans [20]. Many plant WGDs happened 10-80 million years ago[21]. The salmonoid-lineage tetraploidy seen within the rainbow trout genome[22] is the first characterized vertebrate ancient polyploid comparable in time to those in plants. [It should be noted that the trout lineage tetraploidy is thought to be *auto*tetraploid while maize and *Brassica* lineages—those plants studied for fractionation mechanism-- are both *allo*tetraploid.] Trout homeologs have a modal Ks (calculated synonymous nucleotide substitution rate) of 20%, comparable to *Brassica rapa* and maize homeologs, which have a modal Ks of approximately 15%. However, in contrast to these plant genomes, trout fractionation generated many pseudogenes at about the same frequency as totally or partially deleted genes. Assuming that this one autotetraploid reflects "fish", and the two allopolyploids reflect "plants", one simple explanation is that the avidity of the mutational mechanisms underlying purifying selection have adapted to fit effective population sizes, and that, because of pollen, is very large for plants[23] (and this may affect the way plants exhibit "binding site turnover" [24] and see this citation's annotation).

EVOLUTIONARY CONSEQUENCES OF FRACTIONATION

Fractionation is biased because of genome dominance exhibited in ancient allotetraploid genomes. For most ancient plant tetraploidies studied [25], the subgenomes differ in number of ancestral genes surviving fractionation [26-28]. Those sequenced genomes that do display biased fractionation also display genome dominance, where genes on the subgenome that is most intact also tend to express to higher mRNA levels [3], as first documented in maize [29]. This phenomenon is easily visible in two-gene RNAseq FPKM plots of maize homeologs, using datasets from many different cells/tissues/organs/organ components or inductive conditions (Fig. 2). *Gene dominance* is when the genes on one subgenome in a tetraploid tend to express to higher levels than do the gene on the homeologous subgenome; exceptions abound. (Gene dominance, see dotted lines of Fig. 2, when accumulated for all homeologs, constitutes the argument for the trend of *genome dominance*.) Recent analyses on several genomes with polyploidies generalized the link between biased fractionation and genome dominance for plants; it also established a link between unbiased gene fractionation and genome equivalence, as exhibited by the most recent tetraploidy in the banana lineage [3]. Garsmeur, Schnable and coworkers hypothesized, but did not prove, that the difference was allopolyploidy versus autopolyploidy. The Wendel laboratory found that a cotton polyploid occurring 60 million years ago still displays genome dominance [4]. While the Freeling-Wang collaboration on *Brassica* found siRNA coverage—coverage of transposons near genes-- to preferentially mark the not-dominant subgenome with little regard for homeolog expression ratio [30], the Wendel laboratory's data on cotton similarly implicated siRNA, but did not find transposon involvement [4]. Our working hypothesis: the mechanism by which siRNAs lead to the down regulation of nearby genes, as suggested by Hollister and Gaut [31], is "spreading " RNA-dependent DNA methylation and its reinforcing marks, as recently reviewed [32]. Such position effects have been used to hypothesize function for bulk junk DNA and to solve, hypothetically, the C-value paradox [33].

Genes in different GO categories show differing resistances to fractionation following duplication. Reviews document that genes tend to be retained as pairs following a WGD if they encode transcription factors, ribosomal proteins, proteasome core proteins, components of interactive machines or networks, regulatory proteins, signal transducers and similar [34-36] or are associated with many conserved noncoding sequences (CNSs) [37]. Highly expressed genes also tend to be retained, but this is more the case in *Paramecium* than in plants [38]. In Arabidopsis, genes in GO categories retained as tandem duplications at a high frequency are retained as post-WGD pairs at a low frequency, and *vice versa* [36]; this inverse relationship suggested strongly that the *general* reason for gene fractionation resistance was *not* subfunctionalization. This inverse relationship was predicted by the Gene Balance Hypothesis [39,40]. Data from other eukaryotes supports The Gene Balance Hypothesis (reviews[34,41-43]). Gene content of the recently sequenced genomes of trout [22] and *Brassica rapa* [44] are reported to conform to gene-balance expectations. Genes encoding transcription factors, especially "response to" functions, tend to be CNS-rich [45], indicating an abundance of conserved, potentially *cis*-regulatory information. There are at least two explanations for transcription factor gene retention post-WGD: 1) their products participate in protein-protein-DNA complexes [37] *and 2) the genes themselves present long "promoter" targets for subfunctionalization*. Of the 1224 *B. rapa* homeolog pairs with exactly two CNSs, only 9.2% are subfunctionalized at the DNA sequence level, but this increases to a high of 87% for those doublets with 21-61 CNSs [46]. The 5' region of one of these cis-complicated gene doublets is in the inset of Figure 2. Chettoor and coworkers [47] found that genes expressed in maize pollen were significantly resistant to fractionation, perhaps because of increased purifying selection on haploid gametophytes (preferentially applicable to outcrossing species). Data from human tandem duplications indicates that balance in the absence of subfunctionalization may be an important mechanism for retention [48].

Conversely, Duarte and coworkers [49] found that some plant gene families were preferentially fractionated down to singletons. de Smet and coworkers [50] found that many ancestral genes in 20 diverse plants remained mostly single copy (about 2000) or always single copy (a few hundred) even after several rounds of tetraploidies. Overrepresented categories of singleton genes are in DNA metabolism, replication, recombination and DNA repair; one of de Smet and coworkers suggestions was that singletons avoid dominant (like antimorphic) mutations that might disrupt wild type gene function. (Our "dominant mutation" suggestion: cellular processes requiring RNA-DNA or DNA-DNA "loops" [51] with a heteroduplex component might be selected to avoid mismatches). Post-polyploidy changes in gene content in the small genome of the bladderwort, *Utricularia gibba*, also favor singletons [52]; the authors suggest, but do not prove, that the fractionation mechanism is particularly avid in this C-value-decreasing lineage.

Balanced Gene Drive and Subfunctionalization both drive evolution in nonrandom directions. The word "drive" is used here rigorously, as in "meiotic drive" [53,54]. Each WGD *drives* genes with interactive products into the "population"[36] and subfunctionalization *drives* genes with many cis-sites into the "retained" category as well. WGDs—gross sorts of mutations-- cause these drives. The direction of these drives is toward regulatory complexity and redundancy. The rise in morphological complexity (but not other complexities) in green plants has been explained based on balanced gene drive and duplicate gene networks [55], but this hypothesis has not been tested. We hope this bit of mutationist (not selectionist) theory is inoffensive. Note that there is no desire to "drive" to any particular place; there is no "adaptionist paradigm" here. Goldschmidt would probably have called a WGD a "systemic mutation", and argued that the behavior of different sorts of mutations-- not only recombination, selection and population size--, must have a place in a useful theory of evolution. Goldschmidt's 1952 essay is a must-read [56]. Our reigning theory, called the "Modern Evolutionary Synthesis" of the late 1940s, (T. Dobzahnski and several others) disrespects mutation as an evolutionary force.

Fractionation drives last for tens of millions of years, but perhaps not forever. Schnable and coworkers [57] using plant data, and Gout and Lynch [58] using *Paramecium* and yeast data, support similar conclusions: as homeologs increasingly express themselves to different levels, eventually the less expressed homeolog will be lost; this mitigates *balanced gene drive. Subfunctionalization drive* to accumulate genes with complex promoters may be more difficult to mitigate.

Muller's Lexicon [11]				Innan and Kondrashov's [8] Lexicon	
Dominant/ Recessive	Mutant allele (term)	Molecular behavior (example)	Phenotype selection sees (example)	2 essential *cis* sites on duplicate genes	term
D	Wild-type	Specifies product	Wild-type, duplication relaxes selection	+ + + +	"afunctionalization"
R	Both homeologous *cis* elements knockout	Zero product from either gene in target cell	Assume negative, "death"	+ - + -	No term
R	Single knockout, deletion	Zero product from one mutant gene of pair	Possibly none; mutants could be neutral because selection relaxed	+ + + -	Nonfunctionalization
R	Two knockout, deletions	Two *cis* knockouts in each pair	Possibly none, as above	+ + - - - + + -	Nonfunctionalization arrangement Subfunctionalization arrangement
D	Hypermorph	(Deletion of a suppressor); over-producer. (Blocks looping).	"Triplo-insufficient" and negative OR none OR looks like a gain of function but is not.	+ +^{UP} + +	Nonfunctionalization (Nothing "new" here, just "more")
D	Hypomorph	Too little product; Under-producer. (Blocks looping.)	"Haplo-insufficient" and negative, OR none.	+ +^{DOWN} + +	Nonfunctionalization (nothing new here)
D	Antimorph	Product stops function in *trans*	(Antisense RNA, misfolded protein gums up works), Likely negative.	+ + + + m	No term. Could be something new here.
D	Neomorph	New DNA info; (suspect transposons)	Gain-of-function, negative, cis or coding; potential for positive selection	+ + M + +	Neofunctionalization (definitely something new here)
D	Knockout plus hypermorph (an example)	1 gene of the pair with two *cis* mutants in it	Like hypermorph	- +^{UP} + +	Nonfunctionalization, but "molecular subfunctionalization".
D	Mutant in *trans*	Both homeologs up/down regulated	Possible neomorph in a gene regulating the homeologs being studied	+ + + +	No term. Not likely to be two *cis* mutants

Table 1. Muller's classes of mutants on the same page as Innan and Kondrashov' Xfunctionalization terms for mutant arrangements in a gene duplicate with two essential 5' cis-acting sites. + is a wild-type site; - is a site loss-of-function. The arrows in the antimorph row (Row 7) indicate that the product (m) from the mutant gene down-regulates itself and its homeolog (in *trans*). "M" in Row 8 is new information, like a transposon, inserted into the promoter. It is probably best to use the molecular (e.g. "over-producer") rather than Muller's (e.g. "hypermorphic") term when there is no morphological/physiological mutant phenotype. To understand the behavior of any one pair of homeologs (Fig. 2) requires at least one outgroup control and is often an intellectual challenge that cannot be approached using the Xfunctionalization lexicon.

USING FRACTIONATION AS A GENETICAL TOOL TO ACHIEVE PRECISION IN COMPARATIVE GENOMIC EXPERIMENTS

Fractionation analyses can bring inferred, and sometimes proved, function to otherwise pure associations among ENCODE [59] -like features (e.g. DNase Hypersensitive Sites with footprints, DHSs, [60], or DNA protein binding sites via ChIP-seq). Occupied chromatin does not equal function [61-64]. If an ENCODE-like feature existed in the Arabidopsis segment of Figure 2—say a DHS with one protected motif footprint -- and this footprint was fractionated *along with a specific expression character* in a homeologous pair, then that ENCODE-like signature of function is now inferred to have actually functioned as part of a gene.

CNSs, because they are unexpectedly conserved, correlate with past function, but not necessarily in *cis* on the nearest gene. The upper panel of Figure 1 demonstrates how a fractionation pattern (here in a *Brassica*) sometimes allows researchers to infer which gene is the target of any particular CNS activity. For example, since the 5 CNSs located around *At GeneX* (Fig. 1) are deleted in *Brassica* when an ortholog is deleted, they are inferred to act as part of *GeneX*. There are dozens of CNSs spread between Arabidopsis genes *Y* and *Z* in Figure 1. The fractionation pattern in this *Brassica* suggests that none of them act in *GeneY* since its ortholog has been deleted in this *Brassica* and the CNSs remain; *GeneZ* is inferred to include all 47 CNSs. To strengthen specific inferences, experiments on gene fractionation patterns in additional *Brassica* and related radish species – there are several sequenced genomes and all carry the same hexaploidy—can test expectations.

The outcomes of fractionation can be used as part of an analytical method to predict CNS function (‘fractionation mutagenesis’) [20]. Use of this technique requires—at minimum-- a sequenced polyploid with a sequenced outgroup that is not duplicated, each with RNAseq data from many comparable, specific biological endpoints. In plants, only inbred B73 maize (an ancient tetraploid) with a sorghum outgroup (not tetraploid) fits these criteria at this time. Figure 2 shows two-homeolog FPKM plots (from www.qTeller.com) where each point records both FPKMs from one individual RNAseq experiment (like the point for “microspore biological replicate 1); the FPKMs for “comparable” sorghum control endpoints are also indicated. For example, in Panel A, RNA levels in the microspore are off-the-line. Examination of the sorghum microspore data indicates that the maize gene on the x-axis is an over-producer (i.e., a hypermorphic mutant, as opposed to the maize gene on the y axis being an expression knock-down). Our observations indicate that, depending on the biological endpoint, over-producer mutations are approximately as common as under-producer/knockouts. The data in the other three panels, and the legends, support conclusions that are similarly genetic-like in their precision. After proofing FPKM data as reads aligned to the annotated chromosomal segment, examination of the actual DNA sequences resulting from fractionation mutagenesis can deliver candidate “enhancers” or “silencers” for further study.

CONCLUSIONS If there were a tetraploid mammal, it would be the star of the human ENCODE project. The fish tetraploidies could be used intelligently to analyze deeply conserved human ENCODE signatures-of-function.

Not only have studies on polyploid fractionation led to reconsiderations of fundamental evolutionary theory, but fractionation in polyploids permits higher-throughput comparative genomic experiments using ENCODE-like data yielding the logical precision expected of genetic analyses.

ACKNOWLEDGEMENTS
We thank Natalie Todt, M.J.S. Lab , for technical help in obtaining maize and sorghum pollen and microspore RNA sequence available at www.qTeller.com, and to Margaret Woodhouse, M.F. Lab and James Schnable, U. Nebraska, for qTeller activities. Diane Burgess, M.F. lab, provided useful insights and helped with rendering CoGe screenshots publishable. Funded by a USDA NIFA grant #2013-02581 to the authors.

TABLE FOOTNOTES AND FIGURE LEGENDS

Table 1. Muller's classes of mutants on the same page as Innan and Kondrashov' Xfunctionalization terms for mutant arrangements in a gene duplicate with two essential 5' cis-acting sites. + is a wild-type site; - is a site loss-of-function. The arrows in the antimorph row (Row 7) indicate that the product (m) from the mutant gene down-regulates itself and its homeolog (*in trans*). "**M**" in Row 8 is new information, like a transposon, inserted into the promoter. It is probably best to use the molecular (e.g. "over-producer") rather than Muller's (e.g. "hypermorphic") term when there is no morphological/physiological mutant phenotype. To understand the behavior of any one pair of homeologs (Fig. 2) requires an outgroup control and is often an intellectual challenge that cannot be approached using the Xfunctionalization lexicon.

Figure 1. The upper panel is a GEvo graphic (at www.genomevolution.org/coge/) using an eight gene segment of Arabidopsis (*At*) chromosome 4 as query in a blastn sequence comparison to two of its orthologous segments in *Brassica rapa* var. Chiifu (*Br*), the "LF" "dominant" segment and an MF "not dominant" segment. *At-Br*-LF blast HSPs are orange; *At-Br*-MF blast HSPs are brown. The purple rectangles are a compilation of three laboratories' Arabidopsis conserved noncoding sequences, called "VHS-merged" CNSs [51]. "No" means "No fractionation". The inset is a blowup of the indicated 5' region of a crucifer *At GeneZ* with 47 *At-Aethionema* CNSs. "-" indicates deletion, not point mutation[17], of a "+" CNS. (We must assume that, when a gene is deleted, its CNSs are deleted with it.)

Figure 2. Four example categories of mutations recognized during the practice of fractionation mutagenesis in maize plotting RNA levels (FPKM rendered and plotted at www.qTeller.com) of both maize homeologs (ancient tetraploid), with outgroup sorghum (not tetraploid) ortholog RNA-level data embedded in each panel when it exists.. A point is one experiment from the Small Reads Archive; sometimes labels have been condensed, but experiment may be regenerated at qTeller.com. The sorghum expression data is essential to understand the mutants. Slope x/y is *gene dominance*. Ovals enclose focal biologically similar or replicated data points. **A.** *GRMZM2G702426* expresses off-the-line in microspores (haploid male cells in the tetrad) because it is most likely a dominant microspore hypermorph. **B.** Subfunctionalization to extreme cell-within-tissue-component specificity**.** Since there is no sorghum data, it is impossible to know which gene has altered it expression specifically in the adaxial epidermis of the plastochron 7 (ca. 0.7 cm long) leaf primordium, but not in the epidermises of adjacent pre-sheath or pre-blade organ components; RNAseq reads from [65]. **C.** Since both homeologs are vastly over-expressed in pollen expression as compared to the sorghum, we infer a *trans* regulator. **D.** Apparent quantitative subfunctionalization of pollen and microspore expression. Given the sorghum data, this subfunctionalization is expression only, since it is specified by *two mutations in the one gene GRMZM2G134866* comprising an over-production (i.e. dominant hypermorph) in pollen and probably a knockdown of expression in microspores.

FIGURES

Figure 1.

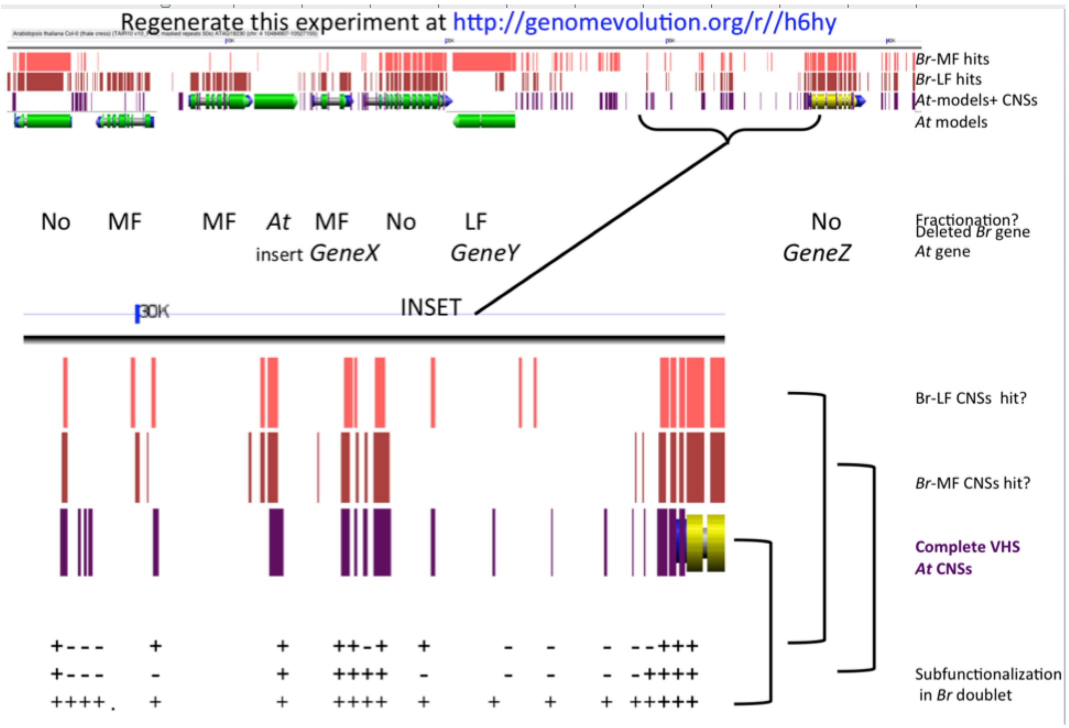
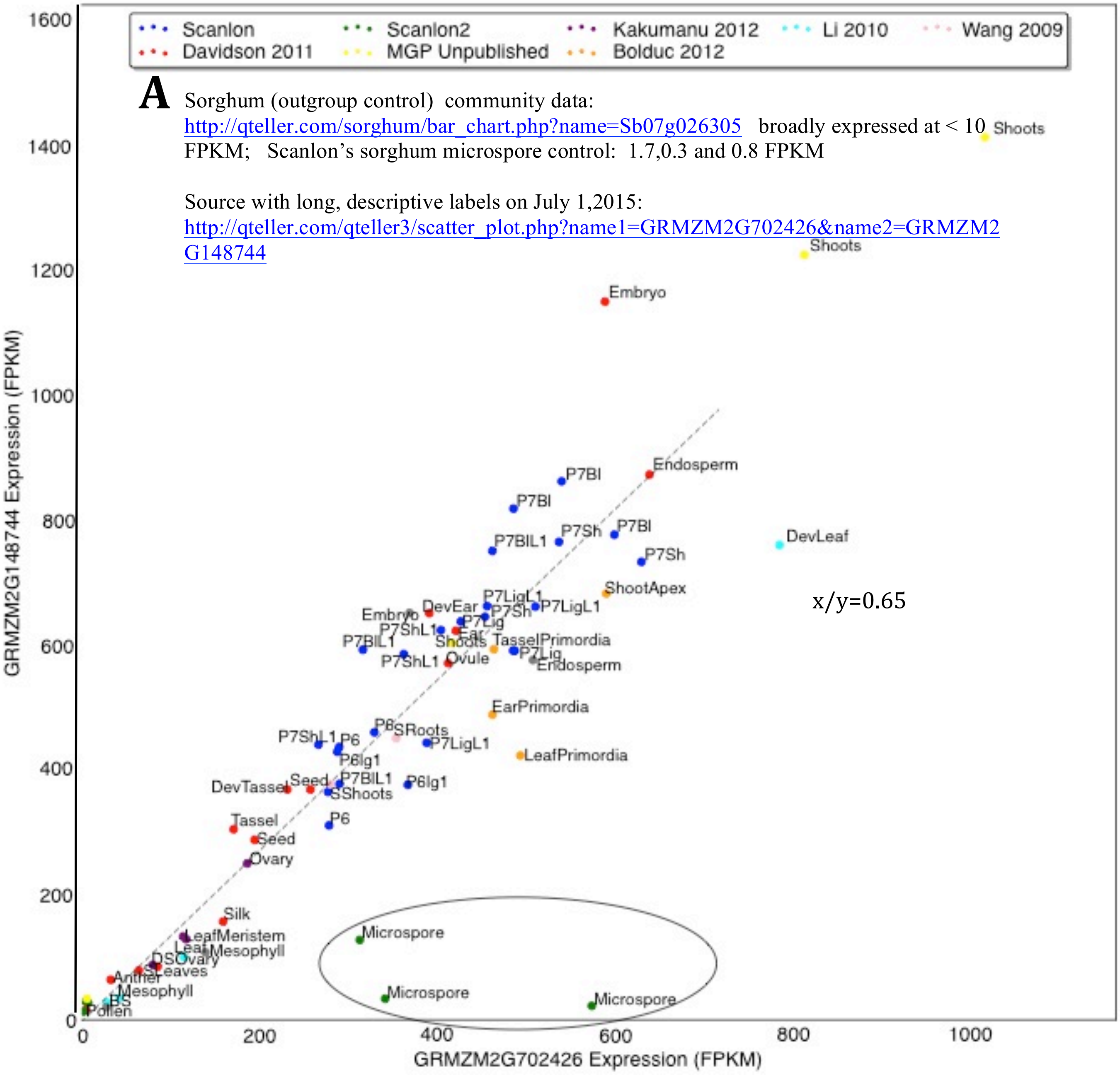


Figure 2 (original Word docs)

Fig.2 A-D on four Word pages. These graphics have the highest resolution.



Sorghum (outgroup control) community data: http://qteller.com/sorghum/bar_chart.php?name=Sb07g026305 broadly expressed at < 10 FPKM; Scanlon's sorghum microspore control: 1.7,0.3 and 0.8 FPKM

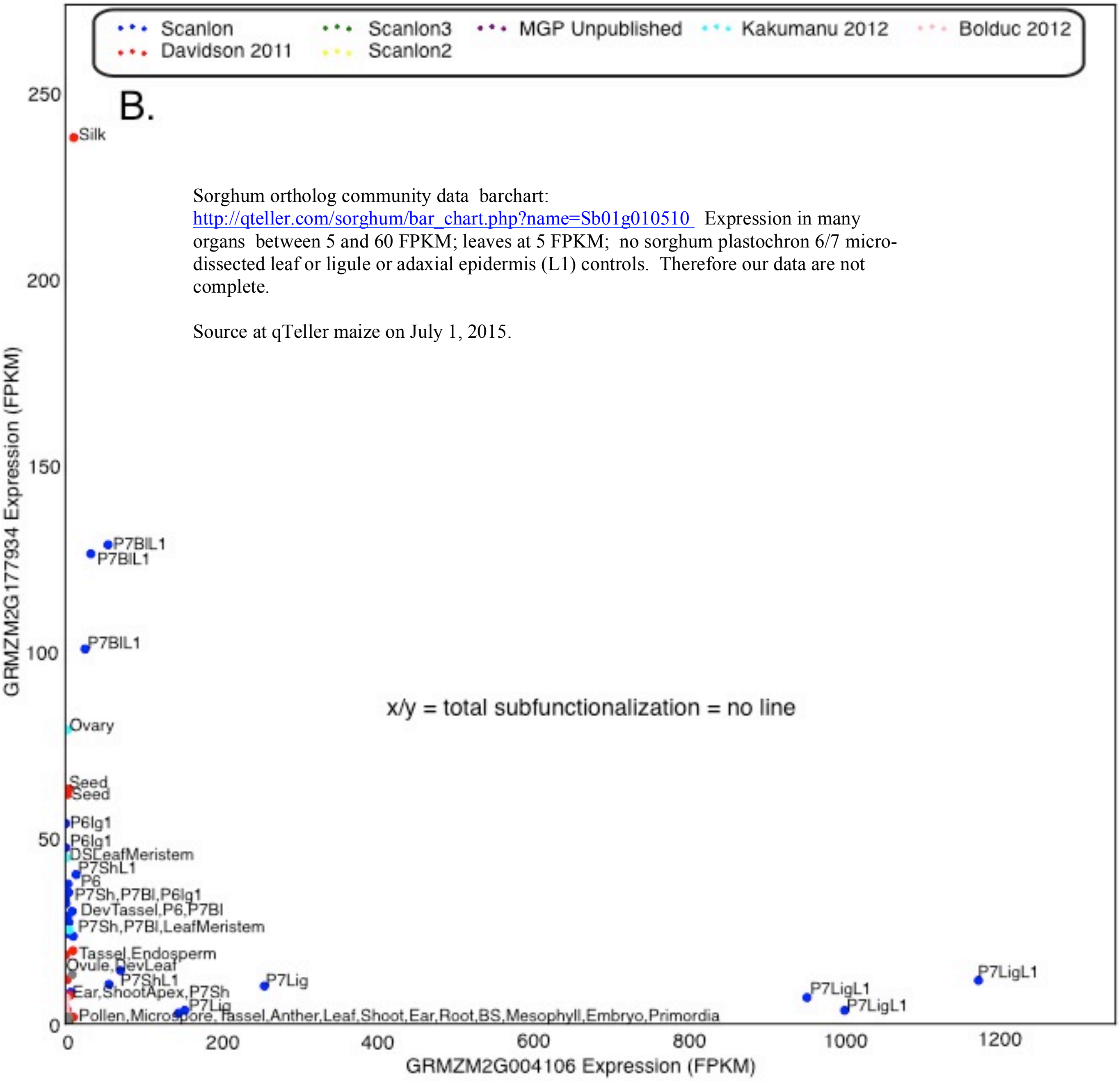
Source with long, descriptive labels on July 1,2015: http://qteller.com/qteller3/scatter_plot.php?name1=GRMZM2G702426&name2=GRMZM2G148744

x/y=0.65

.

.

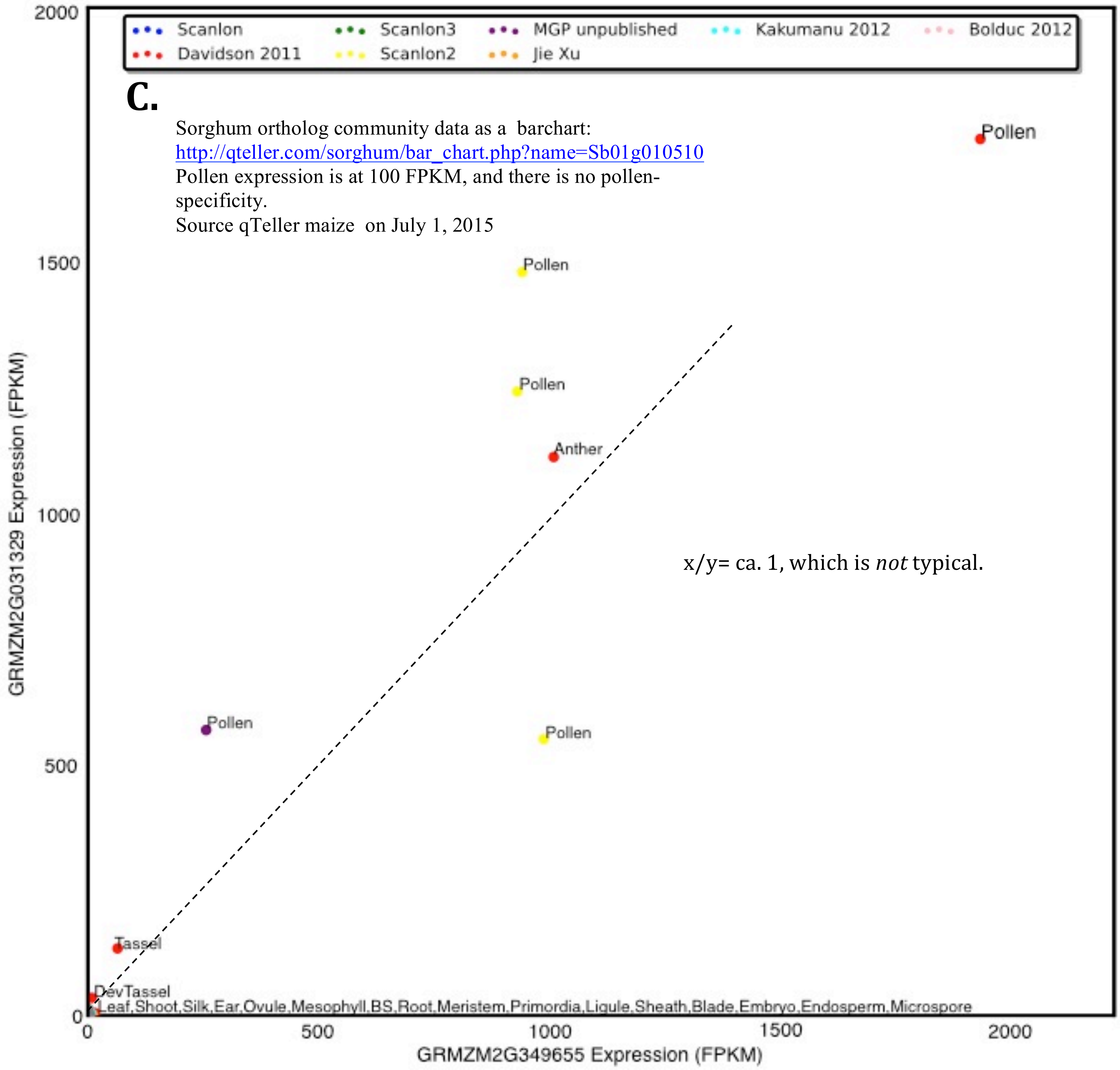
2B



Sorghum ortholog community data barchart: http://qteller.com/sorghum/bar_chart.php?name=Sb01g010510 Expression in many organs between 5 and 60 FPKM; leaves at 5 FPKM; no sorghum plastochron 6/7 micro-dissected leaf or ligule or adaxial epidermis (L1) controls. Therefore our data are not complete.

Source at qTeller maize on July 1, 2015.

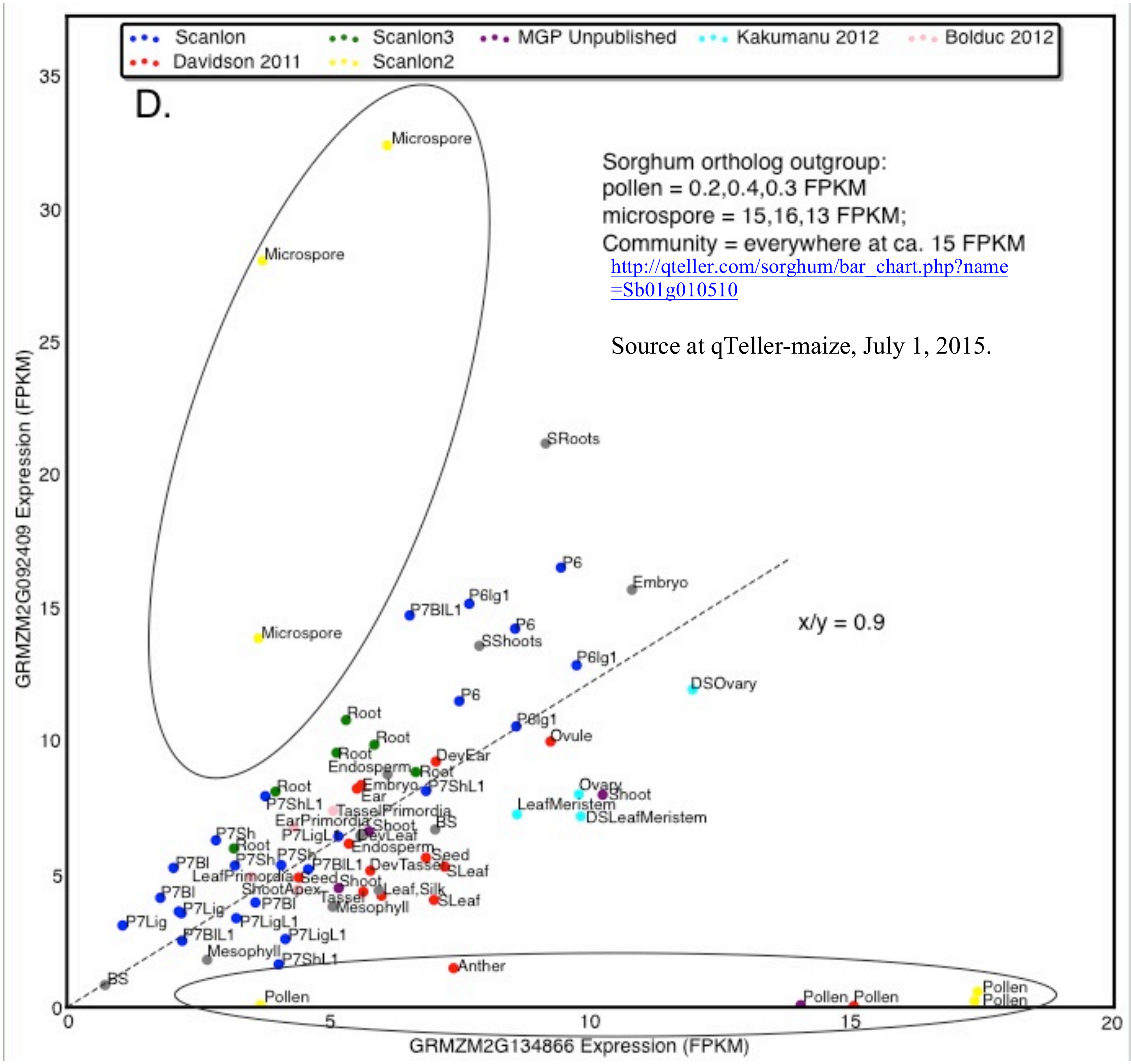
2C



Sorghum ortholog community data as a barchart:
http://qteller.com/sorghum/bar_chart.php?name=Sb01g010510
Pollen expression is at 100 FPKM, and there is no pollen-specificity.
Source qTeller maize on July 1, 2015

x/y= ca. 1, which is *not* typical.

2D



Sorghum ortholog outgroup:
pollen = 0.2,0.4,0.3 FPKM
microspore = 15,16,13 FPKM;
Community = everywhere at ca. 15 FPKM
http://qteller.com/sorghum/bar_chart.php?name=Sb01g010510

Source at qTeller-maize, July 1, 2015.

Stand alone Annotated Reference List for COGD 4.1 MF, JF MJS There are 65 refs.

1. Langham RJ, Walsh J, Dunn M, Ko C, Goff SA, Freeling M: **Genomic duplication, fractionation and the origin of regulatory novelty**. *Genetics* 2004, **166**:935-945.

*2. Murat F, Zhang R, Guizard S, Flores R, Armero A, Pont C, Steinbach D, Quesneville H, Cooke R, Salse J: **Shared subgenome dominance following polyploidization explains grass genome evolutionary plasticity from a seven protochromosome ancestor with 16K protogenes**. *Genome Biol Evol* 2014, **6**:12-33. Rearrangements during post-WGD diploidization respect subgenomes.

3. Garsmeur O, Schnable JC, Almeida A, Jourda C, D'Hont A, Freeling M: **Two evolutionarily distinct classes of paleopolyploidy. *Mol Biol Evol* 2014, **31**:448-454. Ancient WGDs that show biased fractionation also show genome dominance, and those that are evenly fractionated do not display genome dominance; this was judged to be likely the result of allo- versus auto-tetraploidy.

4. Renny-Byfield S, Gong L, Gallagher JP, Wendel JF: **Persistence of Subgenomes in Paleopolyploid Cotton after 60 My of Evolution. *Mol Biol Evol* 2015. Generalizes the phenomenon of genome dominance, first discovered in maize (Ref. 28). Also highlights the invovement of 24bp small RNA (Ref. 30), but perhaps not transposons, as an underlying mechanism.

5. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: **Preservation of duplicate genes by complementary, degenerative mutations**. *Genetics* 1999, **151**:1531-1545.

6. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes**. *Science* 2000, **290**:1151-1155.

7. Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization**. *Genetics* 2000, **154**:459-473.

8. Innan H, Kondrashov F: **The evolution of gene duplications: classifying and distinguishing between models**. *Nat Rev Genet* 2010, **11**:97-108.

9. Lewis EB: **Pseudoallelism and gene evolution**. *Cold Spring Harb. Symp. on Quant. Biol.* 1951, **16**:159-174.

*10. Conant GC, Wolfe KH: **Turning a hobby into a job: how duplicated genes find new functions**. *Nat Rev Genet* 2008, **9**:938-950. An older but accurate introduction to the meanings of the founding terms in the Xfunctionalization vocabulary.

11. Muller HJ: **Further studies on the nature and causes of gene mutations**. *Proceedings of the Sixth international Congress of Genetics* 1932:213-225.

***12. Eichenlaub MP, Ettwiller L: **De novo genesis of enhancers in vertebrates**. *PLoS Biol* 2011, **9**:e1001188. A brilliant use of fractionation in boney fish to identify truly new (neomorphic) enhancers where there was no preceeding enhancer activity. These data are important to the understanding of polyploidy and the evolution of novelty, and are clearly applicable to plant lineages, where WGDs abound.

13. Hughes TE, Langdale JA, Kelly S: **The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. *Genome Res* 2014, **24**:1348-1355. A paper that discovers a high level (13% of pairs) of apparent dominant mutations post tetraploidy in maize leaf homologous parts. The authors use the term "regulatory neofunctionalization", but these mutants are not likely to be anything truly new -- neomorphic mutations-- but rather, the deletion of silencer sites leading to over-expression. Apart from problematic terminology, this is a valuable contribution to our understanding of the high rates of regulatory mutation following WGD.

14. Lee AP, Kerk SY, Tan YY, Brenner S, Venkatesh B: **Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes**. *Mol Biol Evol* 2011, **28**:1205-1215.

15. Woodhouse MR, Schnable JC, Pedersen BS, Lyons E, Lisch D, Subramaniam S, Freeling M: **Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs**. *PLoS Biol* 2010, **8**:e1000409.

16. Tang H, Woodhouse MR, Cheng F, Schnable JC, Pedersen BS, Conant G, Wang X, Freeling M, Pires JC: **Altered patterns of fractionation and exon deletions in Brassica rapa support a two-step model of paleohexaploidy**. *Genetics* 2012, **190**:1563-1574.

*17. Subramaniam S, Wang X, Freeling M, Pires JC: **The fate of Arabidopsis thaliana homeologous CNSs and their motifs in the Paleohexaploid Brassica rapa**. *Genome Biol Evol* 2013, **5**:646-660. As with genes, fractionation of CNSs from whole genome duplicates in *Brassica* is by deletion, not point mutation, 75% of the time. G-box motifs within CNSs are lost by deletion even more often.

18. Zou C, Lehti-Shiu MD, Thibaud-Nissen F, Prakash T, Buell CR, Shiu SH: **Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice**. *Plant Physiol* 2009, **151**:3-15.

19. Wang L, Si W, Yao Y, Tian D, Araki H, Yang S: **Genome-wide survey of pseudogenes in 80 fully re-sequenced Arabidopsis thaliana accessions**. *PLoS One* 2012, **7**:e51769.

20. Freeling M, Woodhouse MR, Subramaniam S, Turco G, Lisch D, Schnable JC: **Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants**. *Curr Opin Plant Biol* 2012, **15**:131-139.

21. Vanneste K, Maere S, Van de Peer Y: **Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution**. *Philos Trans R Soc Lond B Biol Sci* 2014, **369**.

22. Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noel B, Bento P, Da Silva C, Labadie K, Alberti A, et al.: **The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun* 2014, **5**:3657. This WGD is unique among vertebrates because it is recent enough to compare its consequences with those in green plants. Unlike in plants that have been studied, fish frequently fractionate genes via the pseudogene pathway as well as by deletion.

*23. Burgess D, Freeling M: **The most deeply conserved noncoding sequences in plants serve similar functions to those in vertabrates despite large differences in evolutionary rates**. *The Plant Cell* 2014, **26**:1-16. Contains an argument that the major difference between vertebrate and plant CNS detectability deep in evolutionary lineages is that plants (with their gametophytes) have evolved a more avid deletion mechanism that fits their relatively higher effective population sizes.

24. He BZ, Holloway AK, Maerkl SJ, Kreitman M: **Does positive selection drive transcription factor binding site turnover? A test with Drosophila cis-regulatory modules**. *PLoS Genet* 2011, **7**:e1002053. The point of mentioning "binding site turnover" in this cryptic way is to note that sequences deleted cannot turn over. The effective population size of a typical plant population is much greater than that for any *Drosophila* or sea-urchin and very much greater than that for any mammal. *Cis* regultory sequences and genes removed by selection in plants are deleted, not psudogenized. The result is that "binding site turnover" data from

animals, and its companion model -- the "gene regulatory network"-- *may* be less useful in plants and other taxa with larger effective population sizes.

*25. Renny-Byfield S, Wendel JF: **Doubling down on genomes: polyploidy and crop plants**. *Am J Bot* 2014, **101**:1711-1725. Excellent general review on plant polyploidy.

26. Zheng C, Sankoff D: **Practical aliquoting of flowering plant genomes**. *BMC Bioinformatics* 2013, **14 Suppl 15**:S8.

27. Warren R, Sankoff D: **Genome aliquoting revisited**. *J Comput Biol* 2011, **18**:1065-1075.

28. Thomas BC, Pedersen B, Freeling M: **Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes**. *Genome Res* 2006, **16**:934-946.

29. Schnable JC, Springer NM, Freeling M: **Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss**. *Proc Natl Acad Sci U S A* 2011, **108**:4069-4074.

30. Woodhouse MR, Cheng F, Pires JC, Lisch D, Freeling M, Wang X: **Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proc Natl Acad Sci U S A* 2014, **111**:5283-5288. The not-dominant subgenomes of *Brassica* are marked by increased coverage of transposons near genes with 24nt small RNAs, and this asymmetry is true even when the expression of the homeologs is switched. Thus, smRNA coverage marks the not-dominant chromosomes in diagnostic fashion. There may be no way to understand the data in this paper if all transposons are exactly the same with respect to genome dominance.

*31. Hollister JD, Gaut BS: **Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression**. *Genome Res* 2009, **19**:1419-1428. Describes an older, but particularly improtant result from comparing transposons, siRNA targets and RNA levels between *Arabidopsis lyrata* and *thalliana*. In general, the more silencing the lower the expression of genes near transposons. This paper "inspired" the Freeling lab's genome dominance expreriments.

*32. Matzke MA, Mosher RA: **RNA-directed DNA methylation: an epigenetic pathway of increasing complexity**. *Nat Rev Genet* 2014, **15**:394-408. An excellent review on a mechanistically complicated topic.

*33. Freeling M, Xu J, Woodhouse M, Lisch D: **A solution to the C-value paradox and the function of junk DNA: the Genome Balance Hypothesis**, *Mol. Plant 2015, ,* ***8***: 899-910. A unique, testable hypothesis that solves two long-standing problems in eukaryotic biology. As it relates to fractionation, those genes/centromeres with the most junk nearby would tend to be down-regulated/move more slowly to the poles.

*34. Conant GC, Birchler JA, Pires JC: **Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time**. *Curr Opin Plant Biol* 2014, **19**:91-98. An excellent and even-handed review of duplicate gene evolution, and especially how new evidence fits the Gene Balance Hypothesis.

35. Semon M, Wolfe KH: **Consequences of genome duplication**. *Curr Opin Genet Dev* 2007, **17**:505-512.

36. Freeling M: **Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition**. *Annu Rev Plant Biol* 2009, **60**:433-453.

37. Schnable JC, Pedersen BS, Subramaniam S, Freeling M: **Dose-sensitivity, conserved non-coding sequences, and duplicate gene retention through multiple tetraploidies in the grasses**. *Front Plant Sci* 2011, **2**:2.

38. Chen EC, Sankoff D: **Gene expression and fractionation resistance. *BMC Genomics* 2014, **15 Suppl 6**:S19. Highly expressed genes contribute to post-WGD retention in plants, but less so than in *Paramecium.*

39. Birchler JA, Veitia RA: **The gene balance hypothesis: from classical genetics to modern genomics**. *Plant Cell* 2007, **19**:395-402.

40. Birchler JA, Veitia RA: **Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines**. *Proc Natl Acad Sci U S A* 2012, **109**:14746-14753.

41. Edger PP, Pires JC: **Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes**. *Chromosome Res* 2009, **17**:699-717.

42. Birchler JA, Veitia RA: **The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution**. *New Phytol* 2010, **186**:54-62.

*43. Veitia RA, Bottani S, Birchler JA: **Gene dosage effects: nonlinearities, genetic interactions, and dosage compensation**. *Trends Genet* 2013, **29**:385-393. A recent review of gene expression balance and its consequences.

44. Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun JH, Bancroft I, Cheng F, et al.: **The genome of the mesopolyploid crop species Brassica rapa**. *Nat Genet* 2011, **43**:1035-1039.

45. Freeling M, Rapaka L, Lyons E, Pedersen B, Thomas BC: **G-boxes, bigfoot genes, and environmental response: characterization of intragenomic conserved noncoding sequences in Arabidopsis**. *Plant Cell* 2007, **19**:1441-1457.

46. Subramaniam S: **Patterns of computed conserved noncoding sequence loss following paleopolyploidies in the maize and Brassica lineages and their consequences**. In *Plant and Microbial Biology*. Edited by: University of California, Berkeley; 2013:149. [Freeling M (Series Editor), vol Ph.D. .]

47. Chettoor AM, Givan SA, Cole RA, Coker CT, Unger-Wallace E, Vejlupkova Z, Vollbrecht E, Fowler JE, Evans MM: **Discovery of novel transcripts and gametophytic functions via RNA-seq analysis of maize gametophytic transcriptomes. *Genome Biol* 2014, **15**:414. Expression patterns of maize genes with duplicates in subgenome 1 and subgenome 2 indicate that pollen-expressed genes in subgenome 2 (not-dominant) are retained at a higher rate than subgenome 2 genes with other expression patterns. The directly selectable (haploid) genome of the male gametophyte was included in the explanation of this novel, organ-specific sort of fractionation-resistance.

*48. Lan X, Pritchard J: **Long-term survival of duplicate genes despite absense of subfunctionalized expression**. *BioRxiv, a preprint server for biology* 2015, **doi:**http://dx.doi.org/10.1101/019166. A pre-published paper that measures human tandem gene subfunctionalization at many endpoints and finds that balance, not subfunctionalization, seems the best explanation for tandem retention.

49. Duarte JM, Wall PK, Edger PP, Landherr LL, Ma H, Pires JC, Leebens-Mack J, dePamphilis CW: **Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels**. *BMC Evol Biol* 2010, **10**:61.

50. De Smet R, Adams KL, Vandepoele K, Van Montagu MC, Maere S, Van de Peer Y: **Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc Natl Acad Sci U S A* 2013, **110**:2898-2903. This paper generalizes the observation [ref 48] that some sorts of genes end up single copy without regard to polyploidies. They advance a hypothesis: susceptibility to dominant mutation may be selected against.

*51. Burgess DB, Xu J, Freeling M: **Advances in understanding cis regulation of the plant gene with an emphasis on comparative genomics**. *Curr Opin Plant Biol* 2015, **27**:141–147. A review of *cis*-acting regulation of plant genes. The bulk of references are updates from molecular genetics, not comparative genomics, because *cis*-regulation had not been reviewed recently.

52. Ibarra-Laclette E, Lyons E, Hernandez-Guzman G, Perez-Torres CA, Carretero-Paulet L, Chang TH, Lan T, Welch AJ, Juarez MJ, Simpson J, et al.: **Architecture and evolution of a minute plant genome. *Nature* 2013, **498**:94-98. This morphologically reduced, carnivorous plant has about as many genes as a typical core eudicot, but they are surrounded by tiny amounts of noncoding, mostly intergenic DNA. The authors suggest that the fractionation deletion mechanism was accelerated to explain this lowered C-value lineage.

53. Zimmering S, Sandler L, Nicoletti B: **Mechanisms of meiotic drive**. *Annu Rev Genet* 1970, **4**:409-436.

54. Novitski E: **Meiotic drive**. *Science* 1962, **137**:861.

*55. Freeling M, Thomas BC: **Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity**. *Genome Res* 2006, **16**:805-814. A testable but still untested hypothesis where gene and or cis-site retentions, arising as a consequence of WGDs, nudge the direction of morphological (but not other sorts of) evolution into increasing complexity. That mutations might direct evolution, along with selection and chance, is not included within The Modern Synthesis, our current evolutionary theory.

***56. Goldschmidt RB: **Evolution as viewed by one geneticist**. *American Scientist* 1953, **40**:84-98. A must-read layman's essay on a place for mutations in our theory of evolution. This essay radiates honesty about how little we knew (know) about the "intimate archetecture of the chromosome", curiosity as to how evolution really works, and the ideas advanced are still fresh today. Using Goldschmidt's words, a WGD would be a "systemic" mutation engendering "macromutations" necessary (but not sufficient) to the evolution of the founding members of new clades ("hopeful monsters"). Sometimes old is new.

57. Schnable JC, Wang X, Pires JC, Freeling M: **Escape from preferential retention following repeated whole genome duplications in plants**. *Front Plant Sci* 2012, **3**:94.

*58. Gout JF, Lynch M: **Maintenance and Loss of Duplicated Genes by Dosage Subfunctionalization**. *Mol Biol Evol* 2015: **msv095.** A revision of early subfunctionalization theory (Refs. 5-7). As with Ref. 57, helps see how balanced gene drive might terminate if one gene of a pair became expressed below a crucial percent-total threshold, and then be lost.

59. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, et al.: **Defining functional DNA elements in the human genome**. *Proc Natl Acad Sci U S A* 2014, **111**:6131-6138.

60. Spitz F, Furlong EE: **Transcription factors: from enhancer binding to developmental control**. *Nat Rev Genet* 2012, **13**:613-626.

*61. Pfeiffer A, Shi H, Tepperman JM, Zhang Y, Quail PH: **Combinatorial complexity in a transcriptionally centered signaling hub in Arabidopsis**. *Mol Plant* 2014, **7**:1598-1618. A masterpiece of molecular genetics, concluding that unequivocal transcription factor binding to appropriate *cis* sites is a poor indicator of functional binding; much more sophisticated modeling of site-occupancy data is required to avoid simple-minded conclusions.

*62. Biggin MD: **Animal transcription networks as highly connected, quantitative continua**. *Dev Cell* 2011, **21**:611-626. An example of a more sophisticated modeling approach, consistent with Ref. 61 above.

63. Teytelman L, Thurtle DM, Rine J, van Oudenaarden A: **Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins**. *Proc Natl Acad Sci U S A* 2013, **110**:18602-18607.

64. Sullivan AM, Arsovski AA, Lempe J, Bubb KL, Weirauch MT, Sabo PJ, Sandstrom R, Thurman RE, Neph S, Reynolds AP, et al.: **Mapping and dynamics of regulatory DNA and transcription factor networks in A. thaliana**. *Cell Rep* 2014, **8**:2015-2030.

65. Johnston R, Wang M, Sun Q, Sylvester AW, Hake S, Scanlon MJ: **Transcriptomic analyses indicate that maize ligule development recapitulates gene expression patterns that occur during lateral organ initiation**. *Plant Cell* 2014, **26**:4718-4732.