Widespread haploid-biased gene expression in mammalian spermatogenesis associated with frequent selective sweeps and evolutionary conflict

Kunal Bhutani[1,†], Katherine Stansifer[1,†], Simina Ticau[1,†], Lazar Bojic[1], Chloe Villani[2], Joanna Slisz[1], Claudia Cremers[1], Christian Roy[1], Jerry Donovan[1], Brian Fiske[1] and Robin Friedman[*1,†]

[1]Ohana Biosciences, Cambridge, Massachusetts

[2]Center for Cancer Research, Massachusetts General Hospital, Boston, Massachusetts

[†]These authors contributed equally

1 Abstract

Mendel's first law dictates that alleles segregate randomly during meiosis and are distributed to offspring with equal frequency, requiring sperm to be functionally independent of their genetic payload. Developing mammalian spermatids have been thought to accomplish this by freely sharing RNA from virtually all genes through cytoplasmic bridges, equalizing allelic gene expression across different genotypes. Applying single cell RNA

[*]Corresponding author: rfriedman@ohanabio.com

sequencing to developing spermatids, we identify a large class of mammalian genes whose allelic expression ratio is informative of the haploid genotype, which we call genoinformative markers (GIMs). 29% of spermatid-expressed genes in mice and 47% in non-human primates are not uniformly shared, and instead show a confident allelic expression bias of at least 2-fold towards the haploid genotype. This property of GIMs was significantly conserved between individuals and between rodents and primates. Consistent with the interpretation of specific RNA localization resulting in incomplete sharing through cytoplasmic bridges, we observe a strong depletion of GIM transcripts from chromatoid bodies, structures involved in shuttling RNA across cytoplasmic bridges, and an enrichment for 3′ UTR motifs involved in RNA localization. If GIMs are translated and functional in the context of fertility, they would be able to violate Mendel's first law, leading to selective sweeps through a population. Indeed, we show that GIMs are enriched for signatures of positive selection, accounting for dozens of recent mouse, human, and primate selective sweeps. Intense selection at the sperm level risks evolutionary conflict between germline and somatic function, and GIMs show evidence of avoiding this conflict by exhibiting more testis-specific gene expression, paralogs, and isoforms than expression-matched control genes. The widespread existence of GIMs suggests that selective forces acting at the level of individual mammalian sperm are much more frequent than commonly believed.

2 Author's summary

Mendel's first law dictates that alleles are distributed to offspring with equal frequency, requiring sperm carrying different genetics to be functionally equivalent. Despite a small number of known exceptions to this, it is widely believed that sharing of gene products through cytoplasmic bridges erases virtually all differences between haploid sperm. Here, we show that a large class of mammalian genes are not completely shared across these bridges, therefore causing sperm phenotype to correspond partly to haploid genotype. We term these genes "genoinformative markers" (GIMs) and show that their identity tends

to be conserved from rodents to primates. Because some GIMs can link sperm genotype to function, they can be thought of as selfish genetic elements which lead to natural selection between sperm rather than between organisms, a violation of Mendel's first law. We find evidence of this biased inheritance, showing that GIMs are strongly enriched for selective sweeps that spread alleles through mouse and human populations. For genes expressed both in sperm and in somatic tissues, this can cause a conflict because optimizing gene function for sperm may be detrimental to its other functions. We show that there is evolutionary pressure to avoid this conflict, as GIMs are strongly enriched for testis-specific gene expression, testis-specific paralogs, and testis-specific isoforms. Therefore, GIMs and sperm-level natural selection may provide an elegant explanation for the peculiarity of testis gene expression patterns, which are an extreme outlier relative to all other tissues.

3 Introduction

In diploid organisms, Mendel's First Law dictates equal transmission of alleles to the next generation, with strong selective pressure maintaining this 50:50 ratio (Crow 1979). In mammalian spermatogenesis, a long stage of haploid development raises the possibility of allele-biased gene expression and extensive functional variation between mature sperm (Immler 2008). This could be deleterious, for example for important gene products encoded on the X chromosome that would be missing from Y-bearing sperm. However, haploid sperm precursors are equipped with a mechanism for sharing of gene products: cytoplasmic bridges connecting neighboring cells (Braun et al. 1989). Therefore, mature mammalian sperm are thought to be functionally diploid with very rare exceptions.

Most examples of transmission ratio distortion (TRD), i.e. known exceptions to Mendelian inheritance, are attributable to factors other than sperm heterogeneity. However, a handful of sperm functional differences linked to genotype have been reported. The mouse *t* haplotype, a selfish genetic element transmitted at a rate of up to 99% from

heterozygotes, is the best understood case. The mechanism for its TRD is post-meiotic expression and a lack of sharing of *t complex responder* gene products across cytoplasmic bridges, resulting in differential motility (Véron et al. 2009). Likewise, *Spam1* gene products have been shown to be retained in haploid spermatids, underlying TRD in mice carrying certain Robertsonian translocations (Zheng, Deng, and P. Martin-DeLeon 2001). In a mouse model for Niemann-Pick disease, heterozygous knockouts of *Smpd1* have sperm with functional differences in mitochondrial membrane potential associated with their genotype (Butler et al. 2007). Recently, TLR7/8 inhibitors have been reported to differentially affect sperm with the X or Y chromosome (Umehara, Tsujita, and Shimada 2019). Nevertheless, it is widely assumed that most gene products are shared between mammalian gametes, erasing any allelic expression bias.

If, however, sperm functional variation were linked to genotype more often than commonly believed, it might provide an elegant explanation for some peculiar evolutionary phenomena. Testes and spermatids in particular are extreme evolutionary outliers, having far more unique tissue-specific expression patterns, tissue-specific paralogs, alternative isoforms, and selective sweeps compared to other tissues (Kleene 2005). Sexual selection and intragenomic conflict is often invoked to explain this bias, but haploid selection on genes with transmission ratio distortion could easily have contributed (Joseph and Kirkpatrick 2004). For example, alleles with beneficial effects in mature sperm might have deleterious effects in somatic cells, which could drive avoidance of this conflict by evolving sperm-specific paralogs or isoforms. Widespread transmission ratio distortion would be difficult to observe directly due to rapid fixation of beneficial alleles and depletion of deleterious ones, but might leave traces over evolutionary timescales, altering the properties of testis-expressed genes.

TRD enabled by retention of haploid gene products in spermatids would require specific RNA localization rather than free diffusion across cytoplasmic bridges. Recent methodological advances in RNA detection have revealed widespread asymmetric mRNA

distributions in a wide variety of cell types, including up to 70% of mRNAs during *D. melangogaster* development (Lécuyer et al. 2007; Buxbaum, Haimovich, and Singer 2015).

We therefore hypothesized that many endogenous mRNAs would be transcribed in haploid spermatids and incompletely shared across cytoplasmic bridges, resulting in allelic expression bias correlating to the sperm genotype (Fig. 1A). Since mature sperm are transcriptionally and translationally silent, allelic biases in mature sperm protein correlated with the haploid genotype would have to correspond to mRNA expression biases at the haploid spermatid stage. We therefore performed single cell RNA sequencing in spermatids (Fig. 1B) from hybrid mice and cynomolgus macaques, quantifying allele-specific biases in expression. We found surprisingly widespread chromosome-scale biases in single cells allowing confident identification of genes with strong allelic expression links to the genotype, which we term genoinformative markers (GIMs). We show evidence for subcellular localization patterns that help explain their lack of sharing across cytoplasmic bridges, as well as evolutionary consequences consistent with sperm-level natural selection.

4 Results

4.1 Many genes have allelic expression bias reflecting the haploid genotype in spermatids

We first set out to identify cases of incomplete sharing of RNA across cytoplasmic bridges in haploid spermatids (Fig. 1A). This would result in shared information (i.e. correlation) between the allelic expression of a gene and the haploid genotype of the cell, which we call genoinformative expression. Most single cell RNAseq experiments are poorly suited to quantifying allele-specific expression because they do not sequence samples from fully phased individuals, they only sequence a short tag from each RNA molecule (which may not contain a heterozygous site), and they do so with relatively low capture efficiency. To

maximize the accuracy of our allele-specific quantification, we used an F1 hybrid (therefore fully phased) of distantly-related inbred mouse models, C57BL/6 and PWK/PhJ, having over 20 million heterozygous SNPs, compared to roughly 3 million in a human genome (Fig. 1B). We digested testis tissue to isolate single cells from their cytoplasmic bridges, enriched for haploid cells by flow cytometry, and performed full-length single cell RNA sequencing using a slightly modified SmartSeq2 protocol optimized for sensitive RNA capture (Methods).

Of 144 cells obtained from a single male mouse having successful RNA amplification, 126 passed filters as likely singlets with substantial read counts. Principal Components Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) revealed a mixture of three cell types expressing marker genes for spermatids, spermatocytes, and spermatogonia, respectively (Fig. S1A-C). Focusing on the 95 haploid spermatids, we used diffusion mapping (Angerer et al. 2016) to define a pseudotime space covering their differentiation process. The pseudotime ranges from early round spermatids up until the point that the number of genes expressed decreases rapidly at the elongation stage, when transcription arrests (Fig 1C, Fig. S1D). Late spermatid markers such as *PRM3* increase in expression over this pseudotime, while spermatocyte markers such as *SYCP3* decrease (Fig. 1C).

10,991 genes passed filters for calculation of genoinformative expression, including having at least one heterozygous site and having comparable mean expression of each allele (see Methods). We first focused on autosomes rather than sex chromosomes, because we could use the two alleles as an internal control, yielding an easily quantifiable allelic expression ratio within each cell. Visualizing allelic expression in individual haploid cells, we observed strong biases across large stretches of chromosomes, but no consistent bias in diploid controls (Fig. 2A, S2A). Across all haploid autosomes, there was a significant correlation of allelic ratios between neighboring genes that gradually decreased with chromosomal distance, and this correlation was completely absent in diploid controls (Fig.

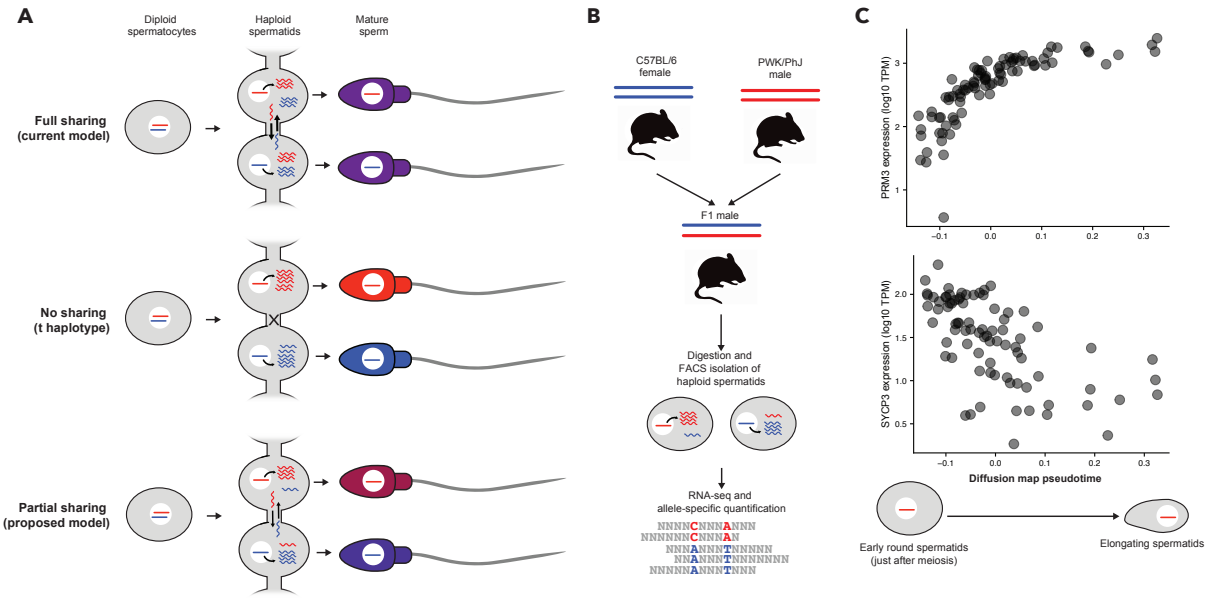
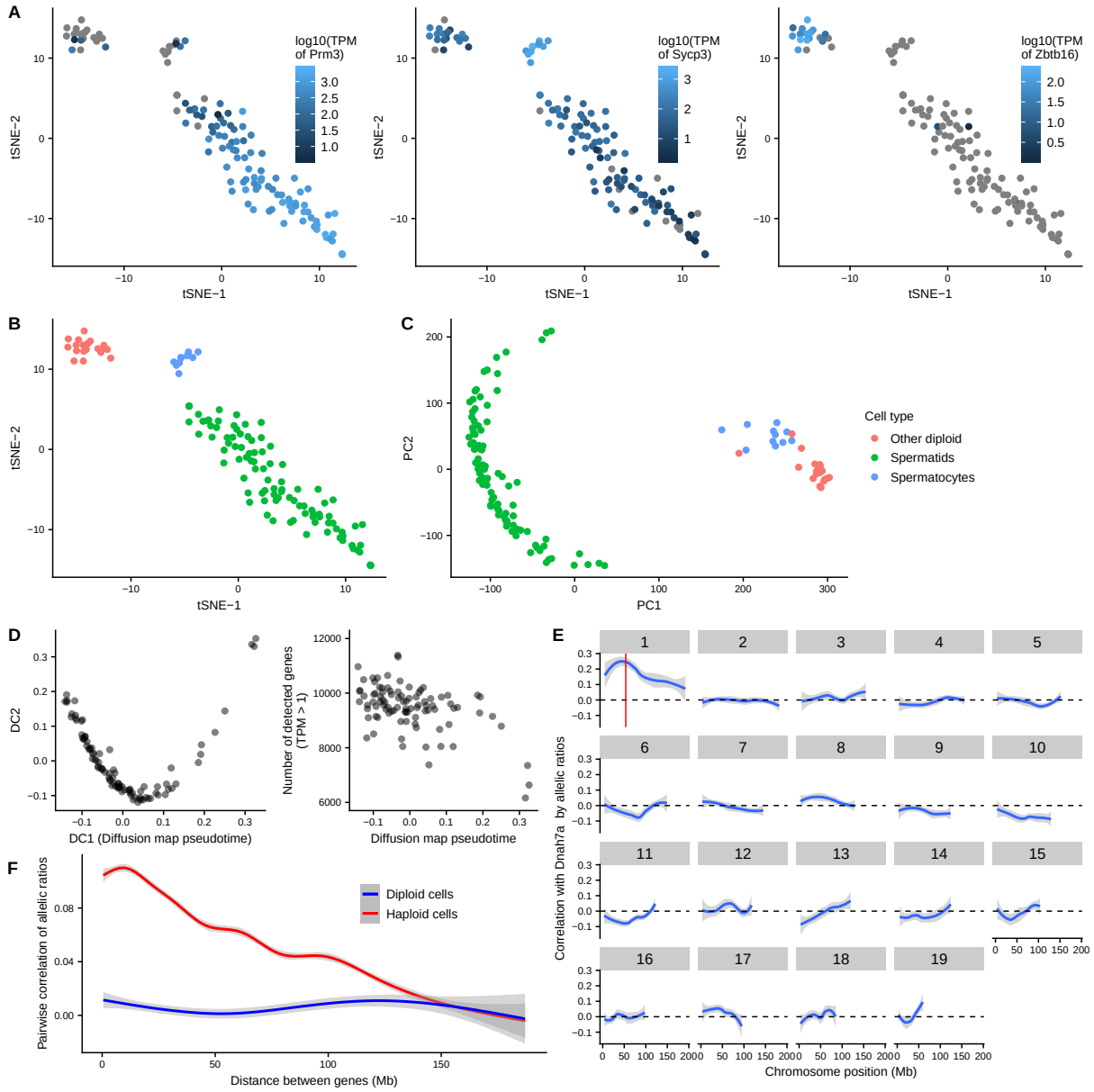


Figure 1: Single cell sequencing of haploid spermatids for assessing allelic bias. **A)** Models for allelic expression bias informative of the haploid genotype (genoinformative expression). The null hypothesis predicts complete sharing between spermatids, erasing any systematic allelic expression differences in mature sperm (top). Selfish genetic elements like the mouse *t haplotype* have virtually no sharing and lead to dramatic allelic differences in mature sperm (center), but incomplete sharing of transcripts would also lead to genoinformative expression (bottom). DNA is represented as straight lines with color representing an allele, and RNA is represented as wavy lines. Sperm color represents the degree of functional links to the allelic genotype. **B)** Experimental setup for single cell RNAseq. We crossed distantly related inbred mouse strains, digested single cells from the testis and enriched for haploid spermatids, and performed full-length RNA-seq and allele-specific quantification. **C)** Pseudotime analysis shows haploid spermatids covered a range from the early round stage (low expression of protamines) to the late elongating phase (very low expression of *SYCP3*)



Supplemental figure 1. Single cell RNAseq of haploid spermatids identifies chromosome-scale correlations in allelic bias. **A)** t-Distributed Stochastic Neighbor Embedding (tSNE) dimensionality reduction for single testis cells enriched for haploid cells. Expression levels in transcripts per million (TPM) are visualized for markers of haploid spermatids (*Prm3*), spermatocytes (*Sycp3*), and spermatogonia (*Zbtb16*). **B)** Cell type annotations based on the above marker genes. **C)** Principal component analysis confirming the tSNE result, showing that all haploid spermatids were strongly distinct from diploid cells. **D)** Left: first two dimensions of diffusion map of haploid spermatids showing the first dimension captured the developmental stage well. Right: Number of genes detected per cell against the first diffusion map dimension (diffusion map pseudotime), showing a decline in those at the latest developmental stage. **E)** Illustration of chromosome-length allelic expression correlation. For one gene on chromosome 1, *Dnah7a* (located at the red line), pairwise correlation of allelic expression ratio was calculated for every gene. Plotted is a loess-smoothed average across each chromosome. Only on chromosome 1 near the *Dnah7a* locus is there a substantial average correlation. **F)** Summary of chromosome-length allelic expression correlations. For each gene, pairwise correlations of allelic expression ratios with all genes on the same chromosome were calculated. The mean correlation in haploid cells or diploid cells across all genes is plotted as a loess-smoothed average. A substantial mean correlation exists for nearby genes in haploid but not diploid cells, and decreases gradually across tens of megabases.

S1E-F). We reasoned that this effect could be explained by a combination of correlation caused by widespread genoinformative expression and degradation of this correlation with distance by recombination. Therefore, we designed a Bayesian probability framework based on an extension of a Hidden Markov Model to infer the haploid genotype of each cell including recombination breakpoints jointly with genoinformativity. Genoinformative expression was modeled as emissions based on the underlying genotype and propensity of an RNA to be shared across cytoplasmic bridges. Intuitively, this model shares information between genes across an entire chromosome for each cell, which means that even weak and noisy genoinformative expression signals in individual genes can aggregate to yield robust signals across large stretches of a chromosome. The model output a probability of genotypes for each cell, and a genoinformativity score for each gene representing the estimated fraction of transcripts retained from its haploid gene expression. Visual inspection confirmed that our inferred genotypes matched the observed expression biases well (Fig. 2A, Fig. S2A). If the inferred genotypes are accurate, the distribution of recombination breakpoints should follow the known recombination density in the mouse genome. Indeed, we saw a significant correlation of inferred recombination density to the published map (Cox et al. 2009) with good agreement at a resolution of 10 to 20 megabases (Fig 2B, S2B-C).

Examining for individual genes the concordance between allelic expression and haploid genotype across cells, we observed a wide range of genoinformativity (Fig. 2C): Many genes, like *Sycp3*, had no association between their allelic expression ratio and the inferred genotype, consistent with our null hypothesis of complete sharing across cytoplasmic bridges erasing allelic expression differences; some, such as *Fer1l5*, had virtually complete concordance with their inferred genotype, suggesting minimal sharing across cytoplasmic bridges; a larger set of genes had clear but intermediate genoinformativity, exemplified by *Ccdc28a*, suggesting partial sharing through cytoplasmic bridges. To determine thresholds for confident genoinformativity, we ran our Bayesian algorithm on shuffled data to create

an empirical background expectation under the null hypothesis of no genoinformative expression (Fig. S2D-E). Thresholds of parameters for both the posterior distribution of the genoinformativity score and the strength of haplotype inference were selected to achieve an empirical False Discovery Rate of 10%. For convenience, genes that met the criteria for confident genoinformative expression were called genoinformative markers (GIMs), regardless of their effect size. Of the 10,991 genes for which we could estimate genoinformativity, 4,354 (39.6%) were confident GIMs and 3,317 (30.2%) were confidently not GIMs (see Methods; Fig. 2D, inset). We were unable to make a confident call for the remaining 3,320 (30.2%) due to marginal signal for genoinformativity. Of the confident genoinformative set, a wide range of effect sizes was seen, but 3,159 (28.8%) had at least a 2-fold average allelic expression ratio in favor of the allele matching the haploid genotype (Fig. 2D).

We were surprised that as many as a third of genes were classified as strong GIMs, so we sought to confirm our assumption that this corresponded to incomplete sharing across cytoplasmic bridges. The chromatoid body is a membraneless organelle (a phase-separated condensate) in germ cells that has been shown to shuttle RNA across cytoplasmic bridges to facilitate sharing (Fig. 2E inset; Ventelä, Toppari, and Parvinen 2003). We found that a published set of genes enriched in the chromatoid body (Meikar et al. 2014) had far lower genoinformativity scores than other genes (Fig. 2E), and that there were fewer GIMs enriched in the chromatoid body than expression-matched controls (Fig. S5C). This confirms that GIMs have different subcellular localization of their RNAs from non-GIMs.

4.2 GIMs have specific subcellular localization resulting in incomplete sharing across cytoplasmic bridges

To identify what mechanisms might be responsible for the differential localization of GIMs, we compared GIMs to non-GIM controls that were matched for expression across spermiogenesis as closely as possible (Fig. S5A, Table S3-4, Methods). Most eukaryotic mRNA

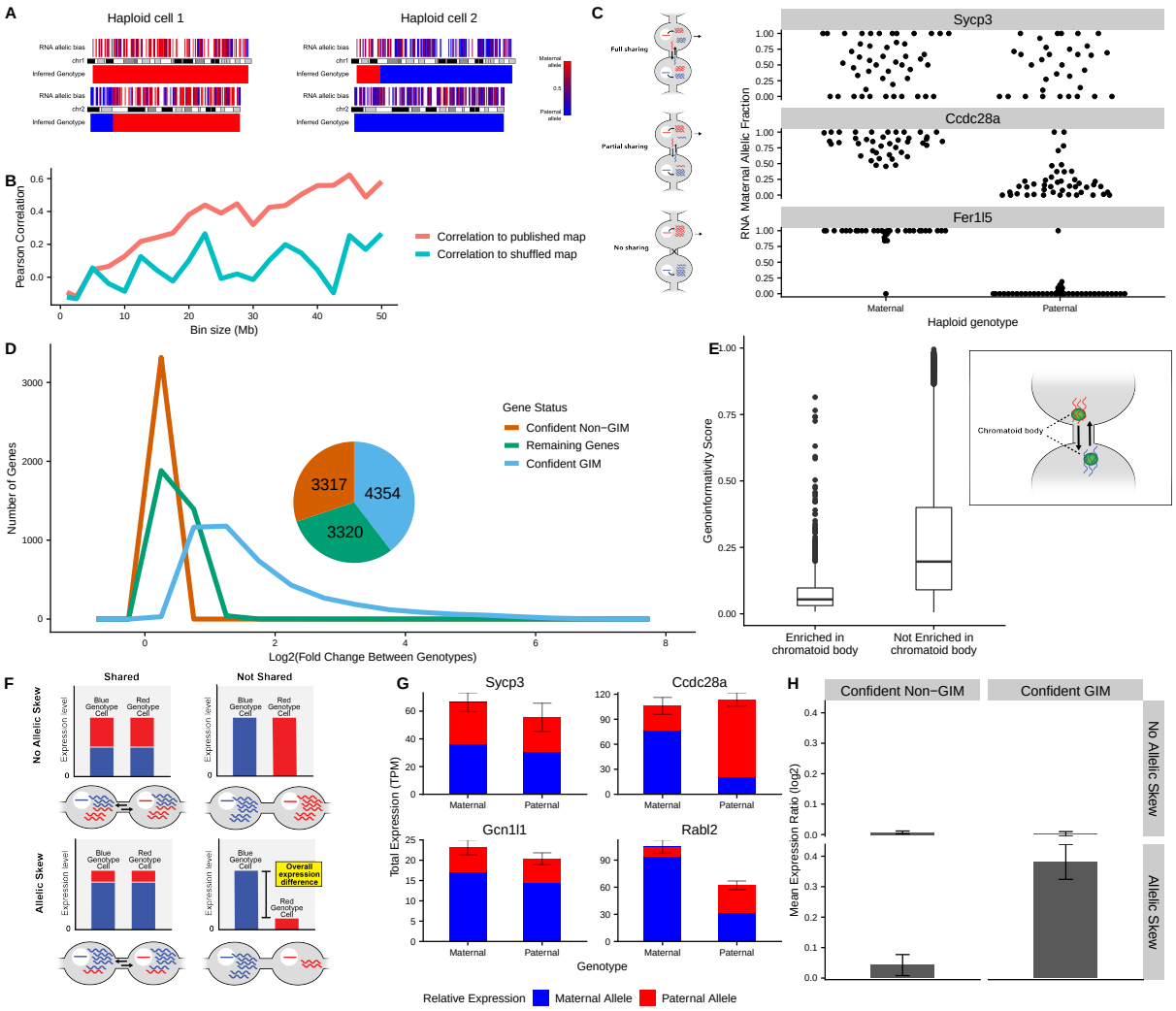
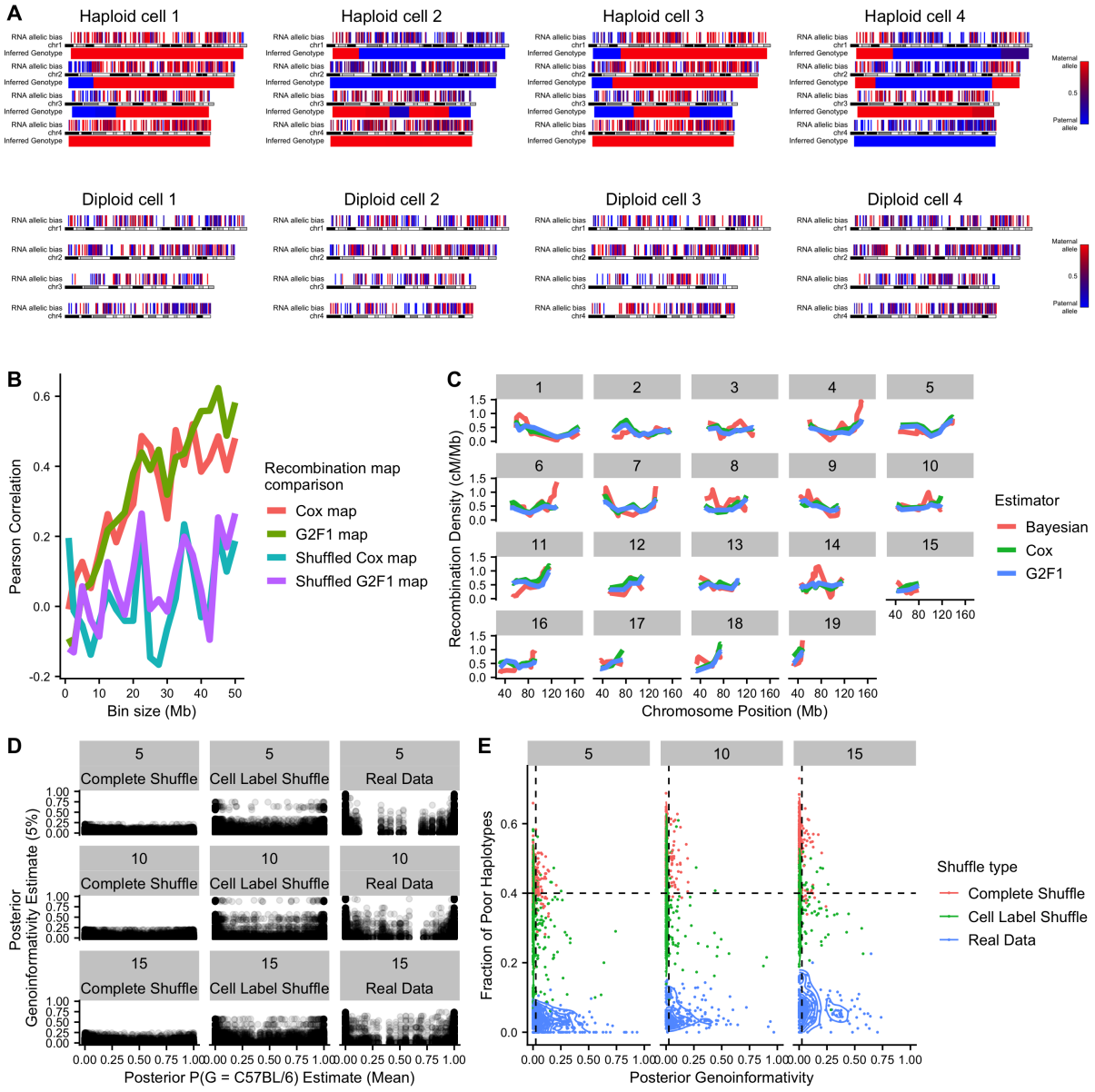


Figure 2: A large fraction of mouse genes exhibit genoinformative expression. **A)** Visualization of allelic bias in the first two chromosomes of two representative haploid cells. Each expressed gene is represented as a vertical line with color representing its allelic ratio (red for more maternal allele, blue for more paternal). Below each chromosome is the genotype inferred by our Bayesian method. **B)** Correlations between inferred recombination densities and a published mouse recombination map (Cox et al. 2009) or a control with recombination densities shuffled between all bins. As bin sizes decrease below about 20 megabases, the variance in our inferred rates increases, causing a degradation of our signal to noise ratio. **C)** Example genes illustrating differing levels of genoinformative expression (right) with their models of sharing (left). *Sycp3* exhibits no association with the haploid genotype, *Ccdc28a* exhibits a strong but incomplete association between the inferred genotype and the expressed allele, and *Fer1l5* exhibits a near-perfect correlation with the inferred genotype. **D)** GIM classification of all genes. Histogram shows the log2 of the expression ratio between the concordant allele (i.e. matching the genotype) over the discordant allele on average across cells. Inset: the total number of genes classified in each category of genoinformative expression. **E)** Genes with mRNAs enriched in the chromatoid body have significantly lower genoinformativity scores. Genoinformativity scores range from zero to one and represent the estimated fraction of transcripts originating from a cell's haploid transcription. Inset: depiction of the chromatoid body's role in shuttling mRNAs across cytoplasmic bridges in haploid spermatids. **F)** A model for how allelic skew (e.g. due to eQTLs) interacts with genoinformative expression. Only genes with both allelic skew and genoinformative expression (not shared) have their mean expression level correlated to the haploid genotype. **G)** Example genes matching the categories in (F). Only *Rabl2* has a significant mean expression difference ($p = 1.5 \times 10^{-5}$, Wilcoxon test). **H)** Summary of expression differences (log2 ratio of genotype concordant with skew to discordant) in all genes in each of the four combinations listed. Only with both allelic skew and GIMs is there an expression difference between cells of differing genotypes.



Supplemental figure 2. Joint inference of genotype and genoinformativity. **A)** Visualization of allelic bias in the first four chromosomes of randomly selected haploid cells and randomly selected diploid cells. Each expressed gene is represented as a vertical line with color representing its allelic ratio (red for more maternal allele, blue for more paternal). Below each chromosome is the genotype automatically inferred by our Bayesian method. **B)** Correlations between inferred recombination densities and two published mouse recombination maps (Cox et al. 2009; Liu et al. 2014) or corresponding controls with recombination densities shuffled between all bins. **C)** Recombination densities across each chromosome (calculated over a 20Mb window) implied by the Bayesian recombination frequencies or for each of the two published recombination maps. **D)** Inferred genotype and genoinformativity for real haploid data and two shuffle types: one permuting both gene and cell labels (complete shuffle) and one permuting only cell labels. Each point is a gene/cell pair, with genotype estimate (x-axis) being a property of the specific gene in a specific cell, and 5% lower bound of genoinformativity (y-axis) being a property of the gene (constant across cells). Three representative chromosomes are plotted (5, 10, and 15). Real data more often have confident genotype estimates and high genoinformativity (upper left and upper right of graph). The cell label shuffle is quite conservative because the genotype structure is maintained, and only the genoinformative expression is randomized. **E)** Summary of the data from (D) illustrating thresholds for calling confident GIMs (dashed lines). Each point is a gene, with poor haplotypes defined as those with less than 95% probability of a genotype. 5% lower bound of posterior genoinformativity probability is plotted on x-axis.

localization is dictated by RNA-binding proteins via sequence motifs in 3′ UTRs (Andreassi and Riccio 2009), so we performed an enrichment analysis for known motifs of RNA-binding proteins that are expressed in spermatids. We identified 26 motifs significantly enriched in GIMs relative to controls, and zero significantly depleted in GIMs (Table S5).

Similarly, a gene ontology enrichment analysis identified strong enrichment for GIMs for specific protein localizations, especially membrane associations and axoneme or other tail localizations (Table S6). To further refine this result, we performed an enrichment analysis with a comprehensive localization database (Binder et al. 2014). This revealed a strong enrichment for genes with annotated localization in neurons, including both dendrites and axons (Table S7), probably reflecting the fact that subcellular RNA localization has been best studied in neurons but is governed by principles applicable across cell types (Ryder and Lerit 2018). Together, these data suggest a mechanism for genoinformativity whereby RNA-binding proteins bring some mRNAs to specific subcellular locations distal from chromatoid bodies, thus partially avoiding sharing across cytoplasmic bridges.

As independent confirmation of our incomplete sharing model for GIMs, we sought to use the much larger set of RNAseq reads that did not overlap a heterozygous site but could be used for estimating overall expression levels. GIMs have allelic expression biases based on the haploid genotype, but because 50% of cells have each genotype, GIMs do not necessarily have a mean allelic expression bias when averaging across many cells (here called allelic skew). However, many genes have a mean allelic skew for other reasons, for example due to expression quantitative trait loci (eQTLs) wherein a genetic variant has differential effects on the expression of a gene. The incomplete sharing model predicts that genes may have different expression levels in spermatids with the paternal versus maternal genotype, but only when they have both an allelic skew and genoinformative expression (Fig. 2F). To illustrate this point, *Sycp3* (Non-GIM, no allelic skew), *Ccdc28a* (GIM, no allelic skew), and *Gcn1l1* (Non-GIM, 2.7-fold allelic skew) all have no difference in

mean total expression from the maternal and paternal genotype cells (Fig. 2G). However, *Rabl2*, which has a 3.0-fold allelic skew and genoinformativity score of 0.45 has a significant difference in expression between the two spermatid genotypes ($p = 1.2 \times 10^{-5}$, t test). Across all genes, we observe that the expression level of GIMs with allelic skew is linked to the haploid genotype in the expected direction, but not for non-GIMs and not for genes without overall allelic skew (Fig. 2H). Therefore both allele-informative and non-allele-informative RNAseq reads support the identity of GIMs and the incomplete sharing model.

4.3 Sex chromosome genes also exhibit genoinformative expression

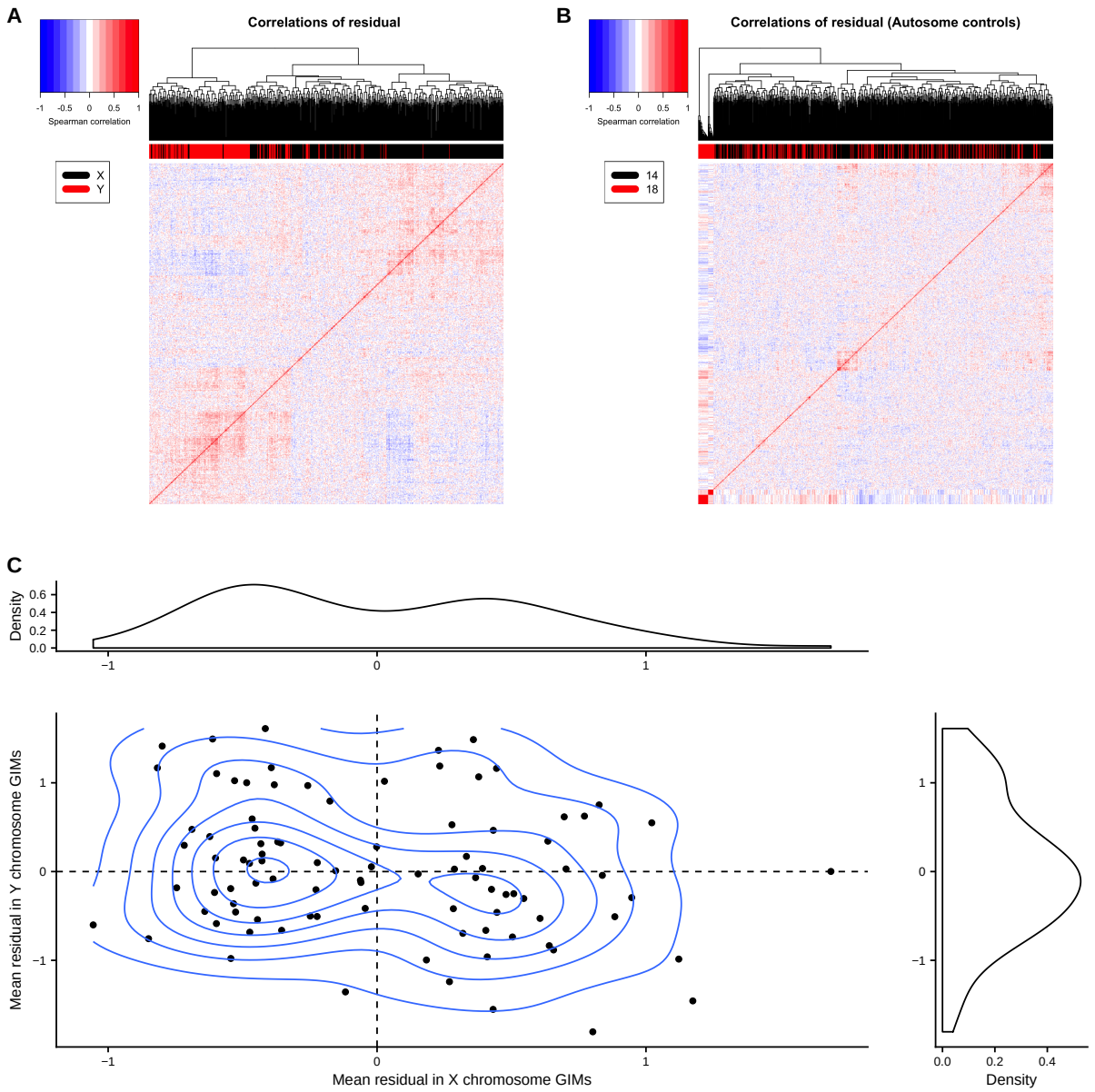
Although our Bayesian method for inferring genotype and genoinformativity cannot be applied to sex chromosomes due to the lack of allelic expression data, genoinformative expression of sex chromosome genes would provide an elegant explanation for models of sex ratio distortion in mice (Cocquet et al. 2012; Eep, Pji, and Ellis Email n.d.). We therefore developed a separate method to identify sex chromosome GIMs based on variation in expression levels rather than in allelic ratios. We started by reasoning that X chromosome GIMs should have correlated expression and be anticorrelated with Y GIMs. Because expression levels in any given spermatid can be strongly influenced by developmental stage, we first corrected for the position in the diffusion map pseudotime. Clustering genes by pairwise correlation after correction, we identified two distinct clusters that corresponded overwhelmingly to the X and Y chromosome, respectively (Fig. S3A). In contrast, performing the same analysis on autosomal controls yielded no similar clusters (Fig. S3B). We selected putative GIMs from these distinct clusters that displayed strong correlation signals (see methods), resulting in 63 X GIMs and 84 Y GIMs (Table S2). Spermatids tend to have high or low mean levels of X GIMs, but not intermediate levels (Fig. S3C). Therefore, sex chromosomes appear to be no exception to the prevalence of

genoinformative expression, at least on a quantitative level.

4.4 Genoinformativity is conserved between individuals and across species

So far, we have only considered mice with one genetic background, so we next asked whether the phenomenon of widespread genoinformative expression extends to other mammals. We dissociated testes from two outbred cynomolgus primates (*Macaca fascicularis*), isolated haploid spermatids and performed single cell RNAseq. Cynomolgus monkeys have the advantage of being highly heterozygous, with ∼13 million heterozygous SNPs per individual, compared to ∼3 million for humans. Because our method for inferring genotypes relies on sharing information across entire chromosomes, we required fully phased chromosomes to quantify genoinformative expression. We therefore combined two phasing methods: a dense, short-range phasing using linked read sequencing, and a sparse, long-range phasing using whole genome sequencing of single haploid spermatids (Fig. 3A). Combining the two sources of information led to densely phased chromosomes for each individual, resulting in 11,654,918 and 10,131,178 phased sites in Cynomolgus 1 and 2, respectively (Fig. S4A).

We were able to quantify allelic expression of a smaller number of genes for cynomolgus spermatids than for mice (7,590 and 4,557 for the two cynomolgus compared to 10,991 in mice), mostly due a smaller number of heterozygous sites. Nevertheless, we observed comparable quality of our genotype inference, including significant correlation of inferred recombination rates between individuals, an expression skew in GIMs with allelic expression skew, and substantial differences between real and shuffled data (Fig. S4B-E). Again using an empirical false discovery rate of 10% in each individual, we classified 50.3% and 52.3% of spermatid-expressed genes as confident GIMs, respectively (Fig. 3B). The effect sizes were comparable to those seen in mice, with 44.6% and 43.3% of spermatid-expressed genes having at least a 2-fold average expression difference between alleles in favor of the



Supplemental figure 3. Sex chromosome GIMs. **A)** Heatmap of pairwise correlations of sex chromosome genes. Correcting for developmental stage (fitting the expression to the diffusion pseudotime position), the residuals of the log expression levels are correlated between all pairs of sex chromosome genes. Two anticorrelated clusters appear, one principally on the X chromosome (black lines above the heatmap), one principally on the Y chromosome (red lines above the heatmap). **B)** Heatmap of pairwise correlations as in (A), but for autosomal control chromosomes with similar numbers of spermatid-expressed genes (chromosomes 14 and 18). No similar broad clusters appear. **C)** Cells have bimodal expression of putative X chromosome GIMs. For each cell, the mean residual log expression across putative X GIMs and Y GIMs is plotted, with density contours. Density plots on the margins show the kernel density of the mean residual for X GIMs (top) and for Y GIMs (right). Most cells have either a high or a low average expression of X chromosome GIMs, but not intermediate. Cells that have high X GIM expression tend to have lower expression of Y GIMs, and vice versa.

haploid genotype. In total, 47.3% of genes that could be quantified met this threshold in either of the two individuals.

Because the two individuals have different heterozygous sites, only 2,366 genes had quantified genoinformativity in both. Among these genes, those that were classified as a confident GIM in one individual had far higher genoinformativity scores in the other individual, and those classified as a confident non-GIM had far lower genoinformativity scores in the other individual ($p < 2.2 \times 10^{-16}$; Fig. 3C). This suggests that within a species, the property of genoinformativity is highly consistent. To look across far larger evolutionary timescales, we compared cynomolgus genes to their orthologs in mouse with a genoinformativity score in each (n = 2,838). Confident GIMs in cynomolgus had higher genoinformativity in mouse than confident non-GIMs ($p < 2.2 \times 10^{-}16$; Fig. 3C), although the relationship was weaker than within a single species. This suggests that the features that confer incomplete sharing across cytoplasmic bridges evolve slowly, so that the identities of GIMs tend to be maintained across evolutionary timescales.

4.5 GIMs show signs of sperm-level natural selection and evolutionary conflict

The substantial fraction of genes having genoinformative expression at the RNA level is surprising, but it does not necessarily imply functional differences in sperm. For example, proteins could be shared across cytoplasmic bridges, nullifying any allelic differences at the RNA level. In contrast, if GIMs lead to functional differences in sperm linked to their genotype, sperm-level natural selection could result in increased evolutionary forces (both purifying and positive selection) acting on GIMs compared to other genes. Given that the identities of GIMs have been maintained across an appreciable evolutionary distance, we reasoned that functional differences in GIMs would lead to detectable signatures in the genome even if they rarely arise. Selective sweeps entail a beneficial allele experiencing positive selection and rapidly reaching fixation in a population, which leaves a signal that

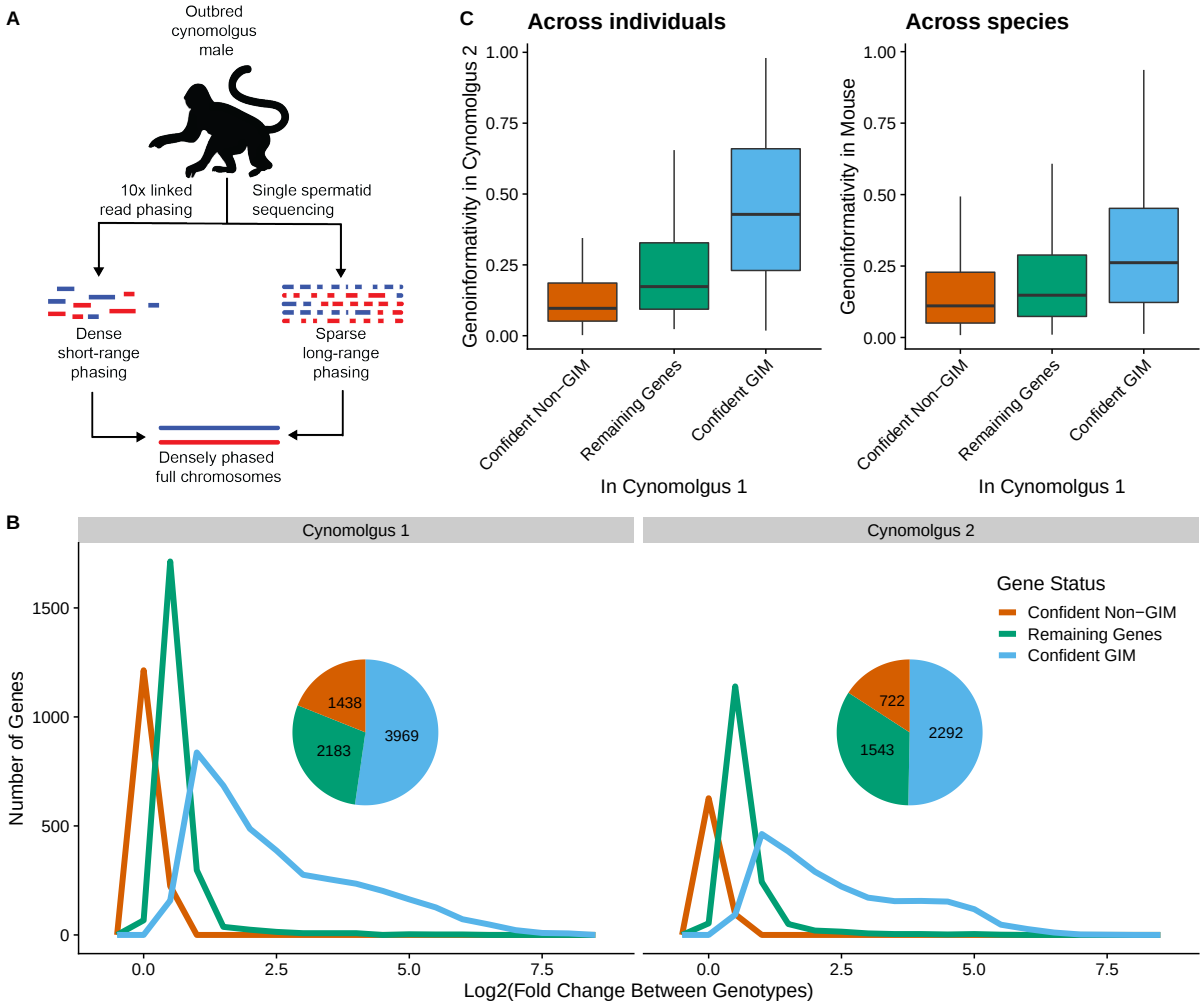
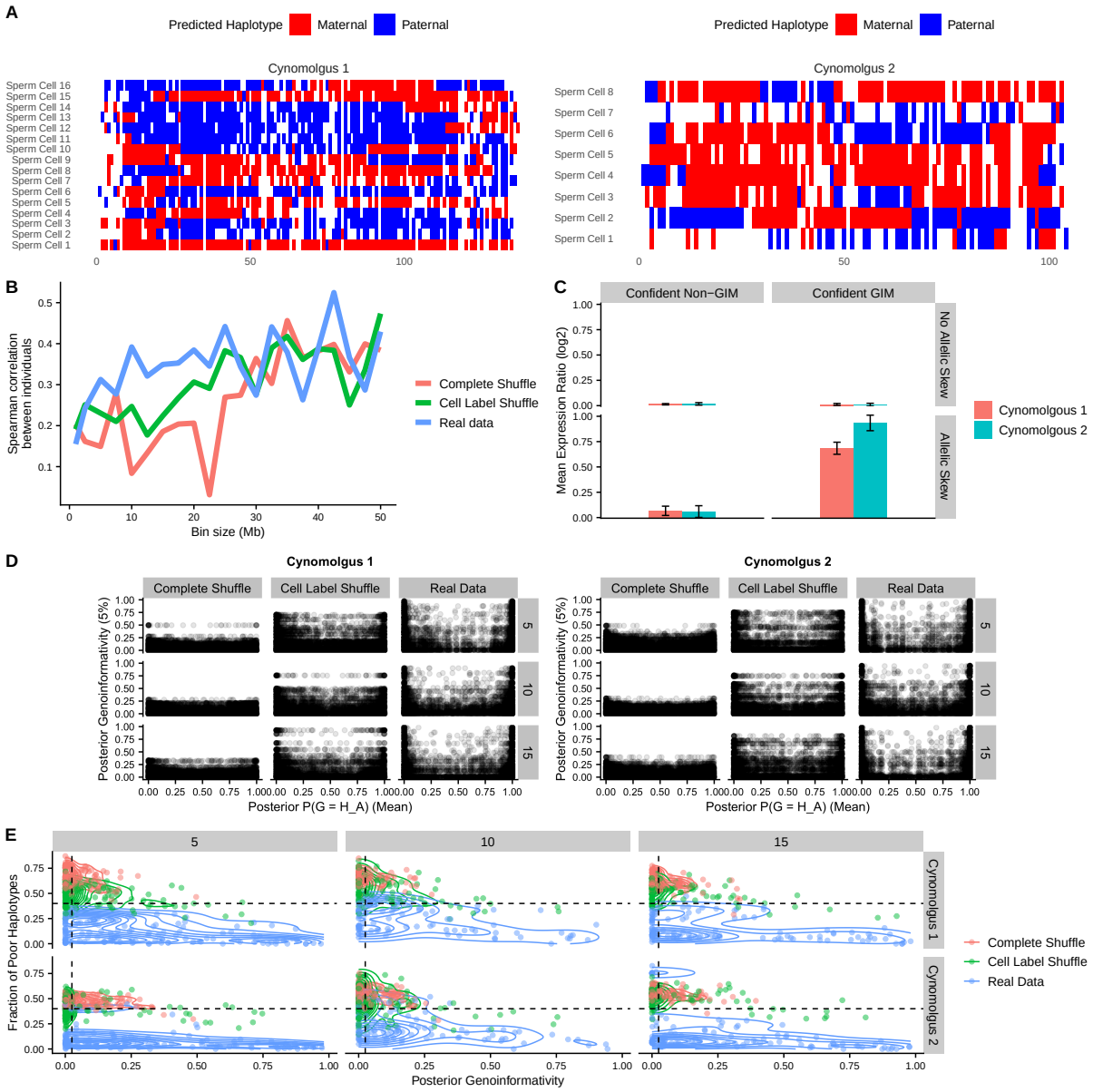


Figure 3: GIMs are conserved between individuals and across species. **A)** Fully phased chromosomes were generated directly from outbred cynomolgus individuals by computationally merging phasing maps from two experimental techniques: short-range phasing from 10x Genomics linked read sequencing, and long-range phasing from whole genome sequencing of several single haploid spermatids. **B)** Genoinformative expression classification of all genes as in Fig. 2D, for each of two cynomolgus individuals. Histogram shows the log2 of the expression ratio between the concordant allele and the discordant allele on average, where the concordant allele matches the inferred genotype. Inset: the total number of genes classified in each category of genoinformative expression. **C)** Conservation of genoinformativity. Genes are categorized based on their genoinformativity classification in Cynomolgus 1 (x axis), and genoinformativity is plotted for these genes in Cynomolgus 2 (left) or orthologs in mouse (right). Genoinformativity scores range from zero to one and reflect the degree of shared information with genotype.



Supplemental figure 4. Cynomolgus primate genotype and genoinformativity inference. **A)** Single cell DNA sequencing data is displayed as phasing blocks called by the 10x Chromium pipeline for chromosome 1. Blocks are assigned to parental chromosomes based on the single cell sequencing data using the algorithm described in the methods section. The resulting patterns show 1-2 recombinations per cell with very few discordant (incorrectly assigned) blocks. **B)** Spearman correlation between recombination densities inferred for the two individuals. Shuffled data showed lower correlations at low to moderate bin sizes. **C)** Summary of expression differences (log2 ratio of genotype concordant with skew to discordant) in all genes in each of the four combinations listed. Only with both allelic skew and GIMs is there an expression difference between cells of differing genotypes, matching the results in mouse. **D)** Inferred genotype and genoinformativity for real haploid data and two shuffle types: one permuting both gene and cell labels (complete shuffle) and one permuting only cell labels. Each point is a gene/cell pair, with genotype estimate (x-axis) being a property of the specific gene in a specific cell, and 5% lower bound of genoinformativity (y-axis) being a property of the gene (constant across cells). Three representative chromosomes are plotted (5, 10, and 15). Real data more often have confident genotype estimates and high genoinformativity (upper left and upper right of graph). The cell label shuffle is quite conservative because the genotype structure is maintained, and only the genoinformative expression is randomized. **E)** Summary of the data from (D) illustrating thresholds for calling confident GIMs (dashed lines). Each point is a gene, with poor haplotypes defined as those with less than 95% probability of a genotype. 5% lower bound of posterior genoinformativity probability is plotted on x-axis.

can be detected by a variety of statistical tests over patterns of variation in the genome. We cross-referenced a set of selective sweeps in wild mouse populations (Staubach et al. 2012) with GIMs and non-GIM controls, either randomly selected from spermatid-expressed genes or matched for expression patterns across spermiogenesis. The GIMs were found in significantly more selective sweep regions than expected by chance ($p = 3 \times 10^{-25}$) corresponding to an excess of 47 ± 4.6 selective sweeps putatively attributable to genoinformativity (Fig 4A, left). Although we do not know of studies of selective sweeps in cynomolgus, we took advantage of abundant predictions of selective sweeps in humans by examining orthologs of cynomolgus GIMs and non-GIMs. Using a set of human selective sweeps (Refoyo-Martínez et al. 2019), we find a significant enrichment of GIMs ($p \leq .013$) corresponding to 9.4 ± 4.2 sweeps putatively attributable to genoinformativity (Fig 4A, right). We corroborated this enrichment for GIMs in a wide variety of tests for selective sweeps in humans and primates on multiple timescales (Fig. S5B). Examining an even larger set of tests for natural selection using 1000 genomes project data (Pybus et al. 2014), we found significant enrichments in a majority of tests (Fig. S5D). Together, this indicates that GIMs are associated with an increased rate of positive selection over evolutionary time.

Sperm-level natural selection poses an evolutionary conundrum: due to its highly specialized function, what is good for the sperm is not necessarily good for the organism. In other words, selection for a beneficial allele in sperm may decrease overall fitness if the allele is deleterious in a somatic cell context (Fig. 4B). Over evolutionary time, this conflict might make genoinformative expression deleterious for genes with somatic functions, but not for genes uniquely expressed in male reproductive tissue. Supporting this hypothesis, we see that GIMs are more likely to be testis-specific in both mouse ($p < 10^{-22}$) and human ($p = 0.006$; Fig. 4C). When it arises, the evolutionary conflict caused by sperm-level selection will cause evolutionary pressure for separating functions for the gene in germ and somatic cells. Examples of this evolutionary pattern include gene

duplication followed by subfunctionalization (Fig. 4B), and testis-specific gene isoforms. As predicted, GIMs are significantly enriched in paralog families that are predominantly testis-expressed in both mouse ($p \leq 6.7\times10^{-12}$; Fig. 4D, left) and human ($p \leq 0.0007$; Fig. 4D, right). Human GIMs are also enriched testis-specific isoforms ($p \leq 1.9 \times 10^{-14}$; Fig. 4E, right), and although we are not aware of similar quality isoform-level mouse datasets, mouse GIMs are significantly more likely to have testis-specific exons ($p \leq 3.7 \times 10^{-9}$; Fig. 4E, left).

Each of these lines of evidence implies that GIMs with these properties are enriched for causing functional differences in sperm, which would require incomplete sharing of proteins across cytoplasmic bridges. In the mouse *t haplotype*, this occurs in part by translating a protein late in spermiogenesis, as cytoplasmic bridges start to break down (Véron et al. 2009). We therefore predicted that GIMs enriched for causing functional differences in sperm would also be enriched in late translation of their proteins compared to other GIMs. Examining a polysome profiling dataset across mouse spermatogenesis (Iguchi, Tobias, and Hecht 2006), mouse GIMs that were functional candidates based on selective sweeps, testis-specific expression, or testis-specific paralogs, were indeed enriched for late translation (p = 0.045, 1.4×10^{-12}, 0.00045, Fisher's exact test; Fig. 4F). However, we did not see enrichment in late translation for GIMs that had testis-specific exons. These results suggest that late translation of GIMs is one mechanism by which they may lead to sperm-level functional differences, causing a higher rate of selective sweeps and avoidance of evolutionary conflict.

5 Discussion

Here we have shown that a large fraction of spermatid-expressed genes are not completely shared between haploid spermatids, resulting in allelic expression that is linked to the haploid genotype, which we call genoinformative expression. Our model for the mechanism for this genoinformative expression is subcellular localization of RNAs, occurring through

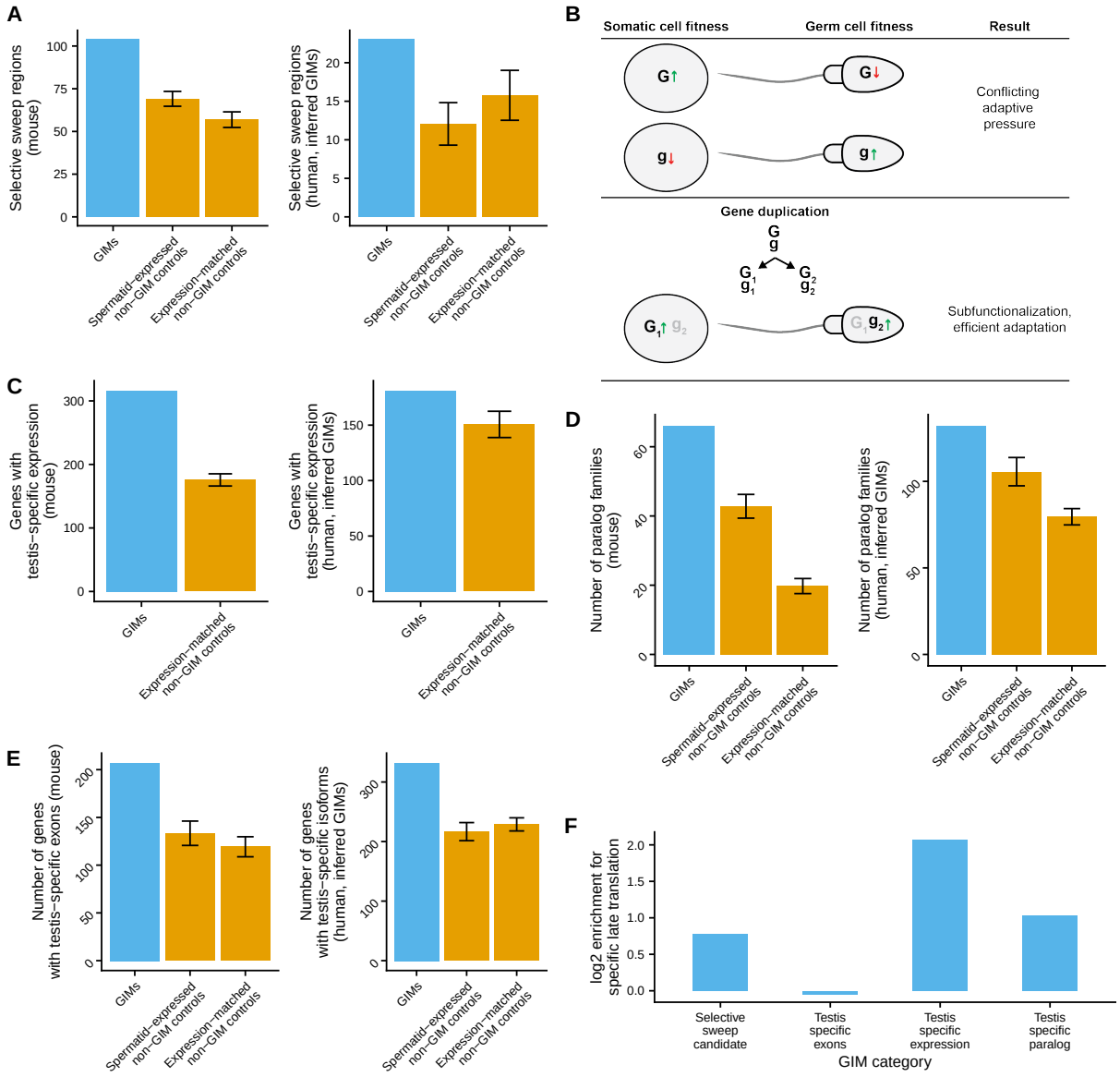
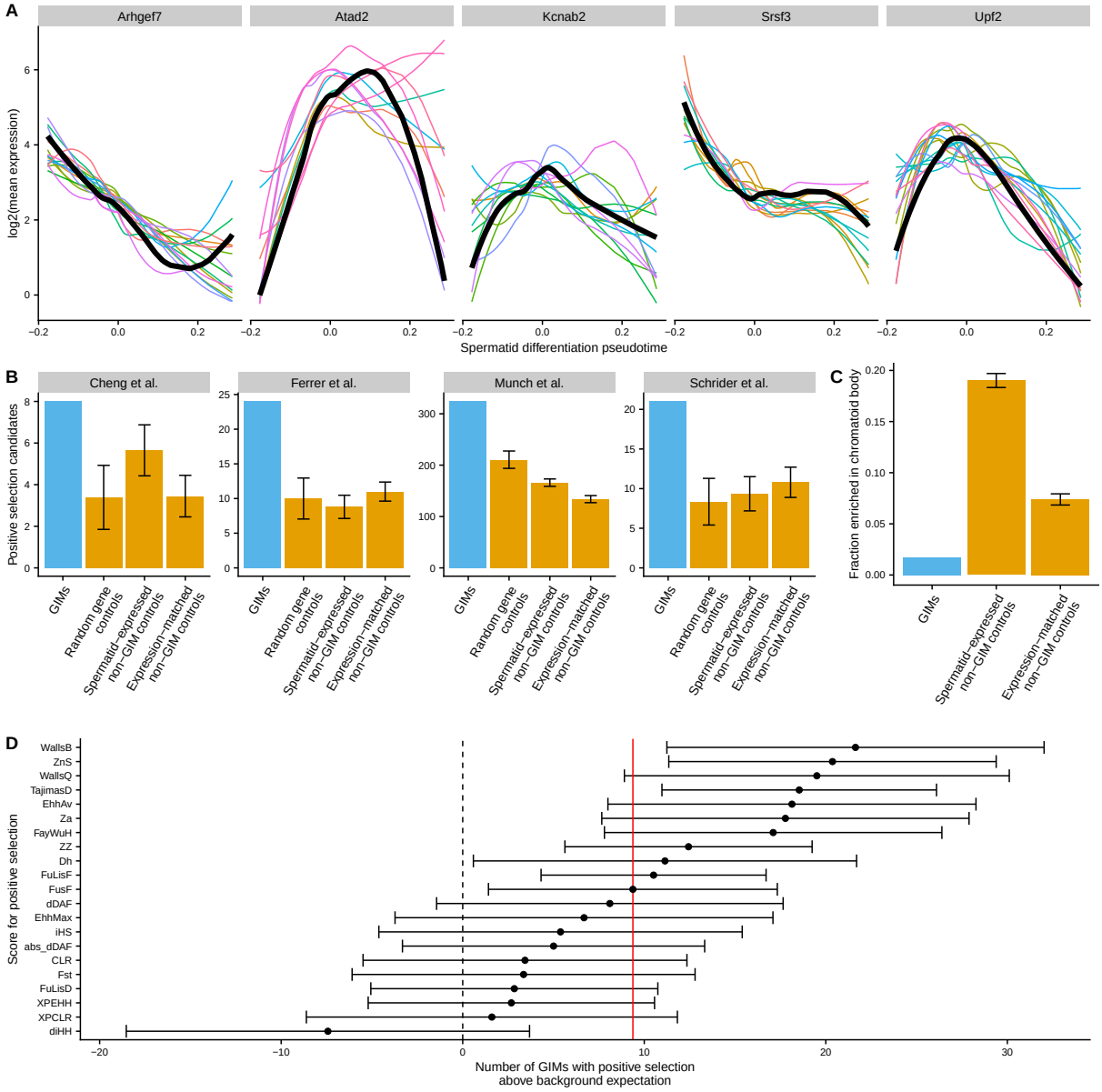


Figure 4: GIMs are associated with sperm-level natural selection and evolutionary conflict. **A)** GIMs are enriched in selective sweep regions in mouse (Staubach et al. 2012) and human (Refoyo-Martínez et al. 2019). Human GIMs were inferred from cynomolgus orthologs. GIMs were compared to control sets (orange bars), either selected from all spermatid-expressed confident non-GIMs, or confident non-GIMs matched to GIMs by their spermatid expression trajectory. **B)** Model for evolutionary conflict between sperm-level and organism-level natural selection. The gene has one allele with beneficial effect in somatic cells but detrimental effect in sperm (G) and one allele with the reverse pattern (g), resulting in positive selection for g at the sperm level, but negative selection at the organism level. A resolution to conflict can be achieved by duplication into two genes, G_1/g_1 expressed in somatic cells and G_2/g_2 expressed in sperm. Selection will then favor the G_1 and g_2 alleles, with no detrimental effects at either level. **C)** GIMs are enriched for testis-specific expression in mice and human, defined as 10-fold higher expression than any other tissue. GIMs were only compared to non-GIMs matched for spermatid expression trajectory, because testis-specific expression is by definition dependent on spermatid expression level. **D)** GIMs represent a higher number of paralog families than non-GIMs in mice and humans. Controls as in (A). **E)** GIMs are enriched in testis-specific isoforms in humans and testis-specific exons in mice. Controls as in (A). **F)** GIMs that are functional candidates are enriched for specific late translation. The GIMs are taken from the blue bars in panel A, C, D, and E, respectively. GIMs in each functional category are compared with GIMs not in that category, and the proportion with specific late translation was calculated. The log2 of the ratio of these proportions is plotted.



Supplemental figure 5. GIM functional characterization. **A)** Illustration of expression-matched control selection for representative GIMs. Thick black lines represent log2 of the loess fit of the expression (in TPM) of GIMs across the spermatid differentiation diffusion pseudotime. Colored lines represent the same loess fit for the 20 genes selected as controls for this gene based on their expression pattern and dropout rate. **B)** The number of positive selection (selective sweep) candidates from several publications (Schrider and Kern 2016; Ferrer-Admetlla et al. 2014; Cheng, Racimo, and Nielsen 2019; Munch et al. 2016) overlapping GIMs or several types of controls. Error bars represent the mean ± standard deviation over the 20 control sets of mock GIMs. GIMs are enriched for selective sweeps in all cases ($p < 0.0276$, $p < 1.01 \times 10^{-6}$, $p < 9.65 \times 10^{-6}$, $p < 8.60 \times 10^{-6}$, respectively). **C)** The fraction of genes overlapping the genes annotated as enriched in the chromatoid body (Meikar et al. 2014) overlapping with each gene category. Bars represent mean ± standard deviation over the 20 control sets of mock GIMs. **D)** Enrichment for GIMs in positive selection candidates based on raw scores for positive selection calculated based on 1000 genomes project data. The background expectation was calculated using the expression-matched non-GIM control set, and error bars represent the mean ± twice the standard deviation of these controls.

RNA-binding protein motifs in the 3′ UTRs or other mechanisms, resulting in depletion of GIMs from the chromatoid body (which facilitates sharing across cytopasmic bridges). GIMs are substantially conserved across populations and evolutionary timescales, so we predict these mechansisms are conserved as well.

In light of this finding, a number of cases of sperm-level functional differences in the literature can be putatively attributed to GIMs (Conway et al. 1994; P. A. Martin-DeLeon et al. 2005; Butler et al. 2007; Véron et al. 2009; Cocquet et al. 2012; Alavioon et al. 2017; Nadeau 2017; Umehara, Tsujita, and Shimada 2019). Despite the growing number of examples of sperm-mediated transmission ratio distortion, it has been widely assumed these are isolated cases and that mammalian sperm are functionally diploid as a rule. The fact that GIMs were so common (over a third of spermatid-expressed genes) surprised us, and suggests that many more cases of sperm-level functional heterogeneity based on genotype will be found.

Mendel's first law dictates that alleles of genes are inherited with equal probability, requiring sperm to be functionally equivalent regardless of their haploid genotype. We believe that remains the case for the majority of genes in mammals at any given time, since transmission ratio distortion has not been commonly observed. However, we show that over evolutionary timescales, GIMs are associated with an increased rate of selective sweeps, suggesting selection at the level of sperm based on functional differences linked to alleles. At first glance, reconciling the sperm-level selection with the predominance of Mendel's first law seems difficult, but there are several reasons to believe they are compatible: 1) We find evidence for only tens to hundreds of selective sweeps across deep timescales and across thousands of GIMs, suggesting that they are relatively rare; 2) Selective sweeps happen quickly on an evolutionary timescale, erasing standing variation and making transmission ratio distortion a rare phenomenon at any one time; 3) Because most GIMs lead to only modest allelic differences (2-4 fold), sperm with these differences may be functionally equivalent or will lead to modest transmission ratio distortion, as is

observed for example in mouse Yq deletions or Slx knockdowns (Conway et al. 1994; Cocquet et al. 2012), which is challenging to quantify in most mammals; and 4) Avoidance of evolutionary conflict by evolving sperm-specific expression removes cases of balancing selection, which might have resulted in observable transmission ratio distortion on standing variation.

While genoinformative expression is widespread at the RNA level, we do not have direct evidence for how common it is at the protein level. One reason to believe there are substantially fewer protein-level GIMs than RNA-level GIMs is that proteins can be shared across cytoplasmic bridges. This is consistent with the fact that GIMs that are preferentially translated late in spermiogenesis, when there is little to no time to be shared across cytoplasmic bridges, are enriched in evidence for selection or avoidance of evolutionary conflict. Even extremely late-expressed GIMs may not always lead to functional differences in sperm, because epididymal exosomes deliver proteins from diploid cells to sperm after they cease transcription and translation, potentially masking allelic differences in mature sperm. Another mechanism for masking the functional consequences of GIMs may be the abundant post-translational regulation of mature sperm, for example during capacitation, which might create larger cell-to-cell variation among sperm than GIMs.

Nevertheless, the ability of GIMs to lead to sperm-level natural selection may have profound evolutionary consequences. We have shown strong enrichments of GIMs for testis-specific expression, testis-specific paralogs, and testis-specific isoforms or exon usage. There are two forces that could give rise to these results: first, evolutionary conflict arises repeatedly in GIMs, which provides an evolutionary advantage to evolve distinct sperm-level function; second, that evolutionary conflict provides pressure to decrease genoinformativity (i.e. increase sharing across cytoplasmic bridges), so that the remaining GIMs preferentially have more sperm-specific expression. It is impossible to distinguish between these models with the data here, but it is likely that both forces contribute.

More comprehensive catalogs of GIMs across species may be necessary to identify which is predominant.

This provokes a profound question: why, from an evolutionary perspective, do GIMs exist at all? For sex chromosome genes, such as *Akap4* (an X-chromosome gene required for sperm motility), it is clear that some degree of sharing is required for sperm function and specific mechanisms have evolved to facilitate sharing (Morales et al. 2002). However, it is not clear that genes need to be shared equally or that absolute functional equivalence is achieved; in some cases, a 2-fold or 4-fold difference in allelic expression may not have strong enough functional effects to exert evolutionary pressure to fully share transcripts. Also, sex chromosome genes are a special case that are hemizygous in males, so there may be even less pressure to share equally for autosomes. For genes or isoforms that are sperm specific, there could in fact be a benefit to sperm-level selection: an intensification of both purifying and positive selection by adding a selective layer on top of organism-level selection. In these cases, there would be no evolutionary conflict between the two selective layers, so some GIMs could become "selfish elements" whose interests are aligned with the organism: improving sperm function, which in turn increases the number of offspring.

The testis-expressed genome has long been a puzzling outlier, including by far the most tissue-specific gene expression, the most tissue-specific paralogs, and the most rapidly evolving genes. The widespread presence of GIMs raises the possibility that sperm-level selection and resulting evolutionary conflicts are common enough to provide an elegant explanation for these phenomena. If functional and molecular heterogeneity of sperm can be understood in enough detail, it is even possible that it could be exploited to isolate and eliminate sperm carrying severe Mendelian disease genes, reducing the risk of disease transmission across generations, as has been previously suggested (e.g. Butler et al. 2007). Given the rarity of GIM-related selective sweeps, it may be technically challenging to identify and leverage this expanded source of sperm heterogeneity. However, the surprisingly widespread existence of GIMs raises the possibility that a wide variety of

severe diseases could be prevented by means of sperm selection.

6 Methods

6.1 Spermatid isolation and cell sorting

Testes were reduced to a single-cell suspension (breaking apart the intracellular bridges between germ cells in the process), using the two-step digestion protocol of Gaysinskaya et al. 2014. Digestions were performed in 6 ml for mouse, with one whole testis as starting material (tunica albuginea removed); and in 30 ml for non-human primate, with 600mg of diced testis tissue as starting material. First, to disperse the seminiferous tubules, testis tissue was incubated in digestion solution 1: Hanks' Balanced Salt solution (HBSS, Sigma Aldrich), 1 mg/ml collagenase Type I (Worthington Biochemical), and 6 U/ml DNAse I (Sigma Aldrich). Incubation was at 37°C for 10 min with horizontal agitation. Tubules were then allowed to settle and the supernatant (containing somatic cells) was discarded. Digestion solution 2 was then added to reduce the tubules to a single-cell suspension: HBSS, 1 mg/ml collagenase, 6 U/ml DNAse, and 0.05% trypsin (Gibco, 2.5% stock solution). Incubation was for 25 min at 37°C with horizontal agitation; tubules were pipetted every 5 minutes, and an additional 0.025% trypsin was added halfway through the incubation. Successful digestion was confirmed by examining the cell suspension under a light microscope. Digestion was quenched with Newborn Calf Serum (Gibco).

After digestion, the single-cell suspension was filtered through a 100 µm cell strainer and centrifuged for 10 minutes at 500g. The supernatant was discarded, and the cell pellet was gently resuspended at 1-2$\times 10^6$ cells/ml in PBS + 5 mg/ml BSA. Hoechst 33342 was added at 10 µg/ml and cells were incubated for 30 minutes at 37°C. Propidium iodide (PI) was added at 1 µg/ml during the last 5 minutes of incubation. Samples were filtered through a 40 µm mesh immediately before sorting.

Single live spermatids were then sorted into 96-well plates as described below, using a

BD FACS Aria, a Beckman Coulter MoFlo Astrios, or a SONY Synergy SY3200 instrument. Our gating strategy was as follows: Selected for 1n cells (spermatids and sperm) based on Hoechst 33342 fluorescence intensity (with 355 nm excitation and a 448/59 nm bandpass emission filter) (Gaysinskaya et al. 2014); Selected for PI-negative cells to get a live population (PI was measured with 561 nm excitation and a 614/20 nm bandpass emission filter)); Enriched for round spermatids by selecting cells with high forward scatter (Bastos et al. 2005)

6.2 Cynomolgus primates

Adult male cynomolgus monkeys (*Macaca fascicularis*) were used for the non-human primate studies conducted at the University of Kentucky. Monkeys were singly housed in climate-controlled conditions with 12-hour light/dark cycles. Monkeys were provided water ad libitum and fed Teklad Global 20% Protein Primate Diet. Spermatid isolation and sorting was preformed at the University of Kentucky with two male monkeys. Monkeys were euthanized, testes were promptly removed and placed in Hanks' Balanced Salt Solution (HBSS) on ice, prior to proceeding to tissue digestion and subsequent preparation of a single cell suspension for cell sorting. All animal care, procedures, and experiments were based on approved institutional protocols from the University of Kentucky Institutional Animal Care and Use Committee IACUC (protocol #2015-2294).

6.3 Single-cell RNA sequencing

Single cells meant for RNA processing were sorted into 96-well full-skirted Eppendorf plates that were pre-chilled at 4°C and were prefilled with 10µL of lysis buffer consisting of TCL buffer (Qiagen) supplemented with 1% beta-mercaptoethanol. Sorted plates with single-cell lysates were subsequently sealed, vortexed, spun down at 300g at 4°C for 1 minute, immediately placed on dry ice to flash-freeze the lysates, and then moved to -80°C for storage. The Smart-Seq2 protocol was performed on single sorted cells as previously

490 described (1-3), with some modifications described below.

491 **Reverse transcription** Single-cells lysates were thawed on ice for 2 minutes, then
492 centrifuged at 3,000rpm at 4°C for 1 minute. 20µL of Agencourt RNAClean XP SPRI
493 beads (Beckman-Coulter) was added to lysates, mixed slowly, to not introduce bub-
494 bles and subsequently incubated at room temperature for 10 minutes. The 96-well
495 plate was then placed onto a magnet (DynaMag-96 Side Skirted Magnet, Life Tech-
496 nologies) for 5 minutes while covered. The supernatant was removed, and the SPRI
497 beads were washed three times with 100µL of freshly prepared 80% ethanol, careful to
498 avoid loss of beads during the washes. Upon completely removing ethanol after the last
499 wash, SPRI beads were left to dry at room temperature for up to 10 minutes. Beads
500 were resuspended in using 4µL of the following Elution Mix: 0.1µL 10µM RT primer
501 (5′AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-
502 3′, IDT), 1µL 10 mM dNTP (Life Technologies), 0.1µL Recombinant RNase-Inhibitor (40
503 U/µL, Clontech), and 2.8µL nuclease-free water. The plates were sealed and then spun
504 down briefly, 5 seconds max to get up to 150rpm. The samples were denatured at 72°C
505 for 3 minutes and placed immediately on ice afterwards. 7µL of the Reverse Transcription
506 Mix was subsequently added in every well, consisting of: 2µL 5x RT buffer (Thermo Fisher
507 Scientific), 2µL 5 M Betaine (Sigma-Aldrich), 0.09µL 1M MgCl2 (Sigma-Aldrich), 0.1µL
508 100µM TSO (5′- AAGCAGTGGTATCAACGCAGAGTACATrGrG+G-3′, Exiqon), 0.25
509 µL Recombinant RNase-Inhibitor (40 U/µL, Clontech), 0.1µL Maxima H Minus Reverse
510 Transcriptase (200U/µL, Thermo Fisher Scientific), and 2.46µL nuclease-free water. Ev-
511 ery well was mixed with the resuspended beads. Reverse transcription was carried out by
512 incubating the plate at 50°C for 90 minutes, followed by heat inactivation at 85°C for 5
513 minutes.

514 **PCR amplification and cDNA purification** 14µL of PCR Mix was added in each
515 well: 0.05µL 100µM PCR primer (5′- AAGCAGTGGTATCAACGCAGAGT-3′, IDT),

12.5µL 2x KAPA HiFi HotStart ReadyMix (KAPA Biosystems), 1.45µL nuclease-free water. The reaction was carried out with an initial incubation at 98°C for 3 minutes, followed by 22 cycles at (98°C for 15 seconds, 67°C for 20 seconds, and 72°C for 6 minutes) and a final extension at 72°C for 5 minutes. PCR products were purified by mixing them with 20µL (0.8X) of Agencourt AMPureXP SPRI beads (Beckman-Coulter), followed by a 5 minutes incubation period at room temperature. The plate was then placed onto a magnet for 6 minutes prior to removing the supernatant. SPRI beads were washed twice with 100µL of freshly prepared 70% ethanol, carefully to avoid loss of beads during the washes. Upon removing all residual ethanol traces, SPRI beads were left to dry at room temperature for up to 10 minutes. The beads were then resuspended in 20µL of TE buffer (Teknova) and incubated at room temperature for 5 minutes. The plate was placed on the magnet for 5 minutes prior to transferring the supernatant containing the amplified cDNA to a new 96-well plate. This cDNA SPRI clean-up procedure was repeated a second time to remove all residual primer dimers and resuspended in a final volume of 15µL of TE buffer. The concentration of amplified cDNA was measured using the Qubit dsDNA High Sensitivity Assay Kit (Life 7 Technologies/Thermo Fisher Scientific). The cDNA size distribution of few selected wells was assessed on a High-Sensitivity Bioanalyzer Chip (Agilent). Expected single cell cDNA quantification was around 0.5-2 ng/µL with size distribution sharply peaking around 2kb.

Library preparation Library preparation was carried out using the Nextera XT DNA Sample Kit (Illumina) with indexing adapters that allow 96 single cell libraries to be simultaneously sequenced. For each library, the amplified cDNA was normalized to a 0.12-0.20ng/µL concentration range. The tagmentation reaction consisted of mixing 1.25 µL of normalized cDNA with 2.5 µL of Tagmentation DNA Buffer and 1.25 µL of Amplicon Tagment enzyme Mix. The 5 µL reaction was mixed well, spun at 3,000 rpm for 3 minutes, incubated at 55°C for 10 minutes and then immediately placed on ice upon completing this incubation step. The reaction was quenched with 1.25 µL of Neutralize Tagment Buffer

and incubated at room temperature for 10 minutes. The libraries were amplified by adding 3.75 µL of Nextera PCR Master Mix, 2.5 µL of mixed indices (Nextera XT Index Kit). The PCR was carried out at an initial incubation at 72°C for 3 minutes, 95°C for 30 seconds, followed by 12 cycles of (95°C for 10 seconds, 55°C for 30 seconds, 72°C for 1 minute), and a final extension at 72°C for 5 minutes. Following PCR amplification, 2.5 µL of each library were pooled together in a 2.0 mL Eppendorf tube. The pool was mixed with 216 µL (0.9X ratio for 2.5 µl of 96 cells pooled together) of Agencourt AMPureXP SPRI beads (Beckman-Coulter) and incubated at room temperature for 5 minutes. The pool was then placed on a magnet (DynaMag-2, Life Technologies) and incubated for 5 minutes. The supernatant was removed and the SPRI beads were washed twice with 1 mL of freshly prepared 70% ethanol. Upon removing all residual ethanol traces, the SPRI beads were left to dry at room temperature for 10 minutes. The beads were resuspended in 100 µL of TE buffer and incubated at room temperature for 5 minutes. The tube was then placed back on the magnet for 3 minutes prior to transferring the supernatant to a new 1.5 mL Eppendorf tube. This SPRI clean-up procedure of the library was repeated a second time to remove all residual primer dimers, using the same approach and the final resuspension was done in 30 µL of TE buffer. The concentration of the pooled libraries was measured using the Qubit dsDNA High Sensitivity Assay Kit (Life Technologies/Thermo Fisher Scientific), and the library size distribution measured on a High-Sensitivity Bioanalyzer Chip (Agilent). Expected concentration of the pooled libraries was 30-50 ng/µL with size distribution of 300-700 bp.

6.4 Single-cell DNA sequencing

Single haploid cells meant for DNA processing were sorted into 96-well full-skirted Eppendorf plates either in (1) 5 µL of Cell Extraction Buffer (4.8 µL of Extraction Enzyme Dilution Buffer, 0.2 µL Cell Extraction enzyme, New England BioLabs) and processed using the PicoPlex kit, or (2) 5µL of Cell Lysis Reaction Mix (4.9 µL of Cell Lysis Buffer,

0.1 µL Cell Lysis enzyme, Yikon Genomics). Sorted plates with single-cell lysates were subsequently sealed, vortexed, spun down at 300g at 4°C for 1 minute, immediately placed on dry ice to flash-freeze the lysates, and then moved to -80°C for storage. All single-cells plates were thawed on ice for 2 minutes, then centrifuged at 3,000 rpm at 4°C for 1 minute prior to processing.

PicoPlex Amplification For PicoPlex amplification, the plates were incubated at 75°C for 10 minutes, followed by 95°C for 4 minutes and held at 22°C until ready for the next steps. The pre-amplification Master Mix, consisting of 4.8 µL of Pre-Amp Buffer and 0.2 µL of Pre-Amp Enzyme Mix was added to each cell, the reaction was mixed well, spun at 3,000 rpm for 1 minute. The PCR was carried out at an initial incubation at 95°C for 2 minutes, followed by 12 cycles of (95°C for 15 seconds, 15°C for 50 seconds, 25°C for 40 seconds, 35°C for 30 seconds, 65°C for 40 seconds, 75°C for 40 seconds), and a hold at 4°C. Following the pre-amplification reaction, each well was mixed well with the Amplification Master mix, consisting of 25µL of Amplification Buffer, 34.2 µL of nuclease-free water and 0.8 µL of Amplification Enzyme Mix. The reactions were mixed well, spun at 3,000 rpm for 1 minute and incubated at 95°C for 2 minutes, followed by 16 cycles of (95°C for 15 seconds, 65°C for 1 minute, 75°C for 1 minute) and a hold at 4°C. The concentration of each cell was measured using the Qubit dsDNA High Sensitivity Assay Kit (Life Technologies/Thermo Fisher Scientific). Expected concentration of the single cell lysates was 20-50 ng/µL with size distribution of 300-1000 bp.

MALBAC Amplification For MALBAC amplification, the plates were incubated at 50°C for 50 minutes, followed by 80°C for 10 minutes and held at 4°C until ready for the next steps. The pre-amplification Reaction Mix, consisting of 29 µL of Pre-Amp Buffer and 1 µL of Pre-Amp Enzyme Mix was added to each cell, the reaction was mixed well, spun at 3,000 rpm for 1 minute. The PCR was carried out at an initial incubation at 94°C for 3 minutes, followed by 8 cycles of (20°C for 40 seconds, 30°C for 40 seconds,

40°C for 30 seconds, 50°C for 30 seconds, 60°C for 30 seconds, 70°C for 4 minutes, 95°C for 20 seconds, 58°C for 10 seconds), and a hold at 4°C. Following the pre-amplification reaction, each well was mixed well with the Amplification Reaction mix, consisting of 29.2 µL of Amp Buffer and 0.8 µLamp Enzyme Mix. The reactions were mixed well, spun at 3,000 rpm for 1 minute and incubated at 94°C for 30 seconds, followed by 21 cycles of (94°C for 20 seconds, 58°C for 30 seconds, 72°C for 3 minutes) and a hold at 4°C. The concentration of each cell was measured using the Qubit dsDNA High Sensitivity Assay Kit (Life Technologies/Thermo Fisher Scientific). Expected concentration of the single cell lysates was 20-60 ng/µL with size distribution of 300-2000 bp.

6.5 Haplotype Phasing

6.5.1 Mouse

We downloaded the combined VCF of laboratory mouse strains from The Mouse Genome project (Keane et al. 2011) and defined maternal and paternal haplotypes utilizing SNPs unique to either C57BL/6J or PWK/PHJ, respectively. For all analyses, we disregarded indels and only considered SNPs. This resulted in a total of 20,986,995 heterozygous SNPs, which overlapped 28,497 expressed genes in mouse round spermatids.

10X Chromium Alignment and Haplotype Calling We created maternal and paternal haplotypes of the two non-human primates using a combination of 10X Chromium linked read sequencing on diploid cells and sparse single cell DNA sequencing on haploid spermatid cells. We aligned the 10X Chromium reads to the *Macaca fascicularis* genome `Macaca_fascicularis_MacFac_5.0` from Ensembl, herein referred to as Ensembl-MF5-G, by first creating a custom reference using the longranger mkref command , and then running longranger using this reference and default parameters to generate 10X Chromium alignment data.

For Cynomolgus 1, the instrument generated 1,850,208 Gel Beads in Emulsion (GEMs)

and the software mapped 819,440,960 reads for 37.2x average coverage across the genome. This resulted in 361,465 haplotype blocks with N50 length of 1.6 MB. Each block contained an average of 36 SNPs for a total of 12,758,999 heterozygous SNPs. Cynomolgus 2 10x Chromium data featured 1,889,596 GEMs that led to 812,899,614 reads mapping at an average coverage of 37.4X. It had 318,516 blocks with N50 length of 1.8 MB and an average of 40 heterozygous SNPs per block, for a total of 12,744,826 heterozygous SNPs.

Mature Sperm scWGS Alignment and Processing We genotyped the haploid sperm scWGS samples using a custom pipeline. First, the paired-end reads were aligned to Ensembl-MF5 using BWA v0.7.5 (Li and Durbin 2009) using the mem option with default parameters. The resulting bam files were sorted using samtools v1.4.1 sort (Li, Handsaker, et al. 2009) and duplicates were removed using sambamba v0.6.6 (Tarasov et al. 2015). `samtools mpileup` with a bed file of the 10X Chromium identified variant positions calculated the allelic depths per heterozygous site. We then filtered the file to only include allelic depths of variant alleles. For Cynomolgus 1, this resulted in an average of 1.2M heterozygous sites per spermatid sample, for a total overlap of 3.3M sites across the 17 spermatid samples. With the 8 spermatid samples for Cynomolgus 2, we covered 3.5M total sites with an average 1.2M sites per sample at roughly 1X coverage.

Creating Chromosome-Length Haplotype Blocks The final step involved stitching the the haplotype blocks generated by 10X Chromium sequencing into chromosome-length haplotypes using the haploid cell haplotypes as a guide. In the case of no recombination, the stitching is trivial and requires only a single sperm sample. Due to recombination, we used multiple sperm single cell WGS samples, and utilized a dynamic programming framework tuned to minimizing the number of recombination events to assign the chromium blocks to maternal and paternal haplotypes.

6.6 Allele-specific Expression Quantification

StringTie Transcriptome Assembly Due to unavailability of a publicly available testes transcriptome of Cynomolgus, we created a custom cynomolgus transcriptome using single cell RNA-seq samples of round spermatid and elongating spermatid cells from the two individuals. First, we aligned the samples to Ensembl-MF5-G using STAR v2.5.3 with default parameters, and merged and sorted the bams using samtools. This resulted in 3.7 billion total reads aligned across the corpus of 480 samples. We fed the merged bam into StringTie v1.3.3 with default options except for -p 39 to indicate a large number of available threads. We compared the StringTie generated transcriptome to MacFas_5.0, a Ensembl-generated transcriptome of Macaca fascicularis using Cufflinks v.2.2.1 gffcompare, and created a dictionary to map the StringTie annotation ids back to known gene symbols.

RNA-Seq Processing and Alignment To reduce allelic bias in read mapping, we used bcftools consensus to generate masked genomes, in which all bases in heterozygous positions were modified to the IUPAC character N in the reference genomes. We used STAR v2.5.3 to align the round and elongating spermatid single cell RNAseq reads, but created custom STAR genomes with either Ensembl GRCm38 or the previously described StringTie-generated transcriptomes. We utilized STAR options –outFilterMultimapNmax 1 to eliminate multi-mapping reads, –alignSJBoverhangMin 4 to force large overlap between RNA-seq reads and the genome, and –outSAMattributes NH HI NM nM MD XS attributes, and removed duplicated reads using sambamba. featureCounts (Liao, Smyth, and Shi 2014) was used to generate gene transcripts per million (TPM) values with options -s 0 for unstranded reads, -p for paired end reads, and -B to require both ends of the read to be mapped.

Generating Allele-specific Counts To quantify allele specific expression of genes, we first assigned each heterozygous SNP to a gene using the snpEff Cingolani et al. 2012

annotate tool using custom snpEff databases. Then, after splitting the aligned RNA-seq bams into chromosome-specific BAMs, we generated the allele counts for each gene and spermatid sample combination. To avoid double-counting of reads that overlapped multiple sites, each read was only counted once in favor of either allele, and if a read matched variants on both alleles, we tagged it as a discordant read and did not utilize it for further analysis. For mouse, we average 145 allele-specific reads per gene per sample across 11,542 phaseable genes in 95 spermatid samples.

We performed an additional step to quantify allele specific counts in the non-human primate samples. Due to limited coverage across the length of an entire gene, StringTie often splits a single Ensembl gene annotation into multiple gene annotations. As such, we summed reads from separate StringTie genes overlapping known annotations. The resulting allele counts files for the monkeys are a combination of known genes annotated by Ensembl and novel genes identified by StringTie only. For Cynomolgus 1, in 187 spermatid samples, we average 122 reads per gene per sample for 8956 phased genes. For Cynomolgus 2, in 185 spermatid samples, we average 131 reads per gene per sample in 8216 genes.

6.7 Haplotype and Genoinformativity Inference

To study haploid-biased gene expression, we require knowledge of the underlying haplotype. We reasoned that if there was true haploid biased expression, it would be possible to infer the haplotype from the allele specific expression data. As such, we derived a model to perform both haplotype and genoinformativity inference simulataneously. Here, we first describe a model for transcript sharing across a syncytium and then extend it to a probability model for observing allele specific reads from round spermatids in a single cell RNA-seq assay.

Model of Genoinformative Transcripts We begin by describing a simple model for the number of transcripts of a single gene g in a single cell c. The total number of transcripts T in the haploid cell is the combination of external transcripts E and retained transcripts R.

$$T = E + R \tag{1}$$

Here, external transcripts indicates transcripts that were not transcribed by the haploid cell, but rather were transported into the cell through the cytoplasmic bridge. Retained transcripts are the transcripts that were transcribed by the cell and not shared through the cytoplasmic bridge.

We can also write down the total transcripts T as the combination of transcripts from the maternal allele of the gene M or the paternal allele of the gene P.

$$T = M + P \tag{2}$$

Note that we can marginalize the maternal and paternal transcripts in terms of external and retained transcripts.

$$\begin{aligned} M &= E_M + R_M \\ P &= E_P + R_P \\ T &= E_M + E_P + R_M + R_P \end{aligned}$$

Before deriving a model for genoinformativity, we introduce two last definitions in the form of ratios. The ratio of E_M to E, or the skew of transcripts towards the maternal allele S, and the ratio of R to T, or the genoinformativity of the transcript.

$$S = \frac{E_M}{E} \tag{3}$$

$$G = \frac{R}{T} \tag{4}$$

Assuming no eQTL effects, genome imprinting, technical bias, or other mechanisms for differential allelic expression, this allelic skew S is 0.5, i.e. the number of haploid cells that contain maternal and paternal genotype are equal and the number of transcripts transferring into the cell is equal from either allele.

Haploid Cell with Maternal Allele Given the previous system and definitions, we now derive the empirical genoinformativity for a single haploid cell. Consider the case of a cell c having the maternal allele for the gene or haplotype H_M. Then, we further deconvolve the total transcripts by the transcripts from the maternal allele M and the transcripts from the paternal allele P. Note that this classification is only relevant for autosomes, where it is possible to have transcripts from either chromosome in the haploid cell.

$$\begin{aligned} M|H_M &= E_M|H_M + R_M|H_M \\ &= E_M + R_M|H_M \end{aligned} \tag{5}$$

$$\begin{aligned} P|H_M &= E_P|H_M + R_P|H_M \\ &= E_P \end{aligned} \tag{6}$$

Since the cell has a maternal allele only for the gene of interest, there are no retained reads from the paternal allele. Finally, let's express the total transcripts T in terms of the maternal and paternal transcripts.

$$\begin{aligned} T|H_M &= E_M + E_P + R_M \\ &= E + R_M \end{aligned} \tag{7}$$

Given equation 3, 4, and 7, we can restate equation 5 as

$$\begin{aligned} M|H_M = ES + R_M &= (T - R_M)S + R_M \\ &= TS + (1 - S)R_M \\ &= TS + (1 - S)TG \\ &= (S + (1 - S)G)T \end{aligned} \tag{8}$$

We can derive similar equations for $P|H_M$, $M|H_P$ and $P|H_P$.

$$\begin{aligned} P|H_M &= ((1 - S) - (1 - S)G)T \\ M|H_P &= (S - SG)T \\ P|H_P &= ((1 - S) + SG)T \end{aligned} \tag{9}$$

Probability Model for Allele-Specific Reads We now focus our attention on developing a model for observing allele-specific reads using single cell RNA-Seq from haploid round spermatids. We derive a probability model for observing counts of alleles from the maternal allele C^M and paternal allele C^P for N individuals and G genes. Given parameters θ, each cell i and gene j is independent of each other and the collective probability can be written as:

$$P(D|\theta) = \prod_i^N \prod_j^G P(C_{ij}^M, C_{ij}^P|\theta) \tag{10}$$

For simplicity, we will write the set of counts C_{ij}^M and C_{ij}^P as D_{ij} where applicable. The main reason we are able to treat each set of counts independently is because we marginalize the probability over the haplotype H_{ij} of cell i at gene j.

$$P(D_{ij}|H_{ij},\theta) = P(D_{ij}|H_{ij}=H_M,\theta)P(H_{ij}=H_M|\theta) + P(D_{ij}|H_{ij}=H_P,\theta)P(H_{ij}=H_P|\theta) \tag{11}$$

Using the above formulation, it is possible to split the inference goal into two separate sub-tasks: haplotype inference and genoinformativity inference.

Haplotype Inference We use a Markov chain across a single chromosome to estimate the haplotype given a recombination rate r.

$$P(H_{ij}|\theta) = P(H_{ij}|H_{ij-1})P(H_{ij-1}|\theta) \tag{12}$$

where

$$P(H_{ij}|H_{ij-1}) = \begin{cases} (1-r) & H_{ij} = H_{ij-1} \\ r & Hij \neq H_{ij-1} \end{cases} \tag{13}$$

We set the initial probability of each cell's haplotype to be equal at 0.5.

Genoinformativity Inference Given the haplotype H_{ij} of cell i at gene j, the counts of the maternal and paternal allele follows from the generative model described above. Due to overdispersion in RNA-seq data, we model the counts using a beta-binomial distribution, which is specified by shape parameters α and β. In fitting the model, we only fit the shape parameter β and reparameterize α in terms of skew S and genoinformativity G. More explicitly, we can model the system as

$$\frac{\alpha_j}{\alpha_j+\beta_j} = \begin{cases} S_j + G_j(1-S_j) & H_{ij} = H_M \\ S_j - G_j(S_j) & H_{ij} = H_P \end{cases} \tag{14}$$

$$P(D_{ij}|H_{ij},\theta)P(\theta) = P(D_{ij}|H_{ij},S_j,G_j,\beta_j)P(\beta_j)P(G_j)P(S_j) \tag{15}$$

$$P(D_{ij}|H_{ij},S_j,G_j,\beta_j) = \text{Beta-Binomial}(\alpha_j,\beta_j,D_{ij}) \tag{16}$$

In addition to overdispersion, single cell RNA-seq data also contains high amount of allelic dropout and amplification of a single molecule. To alleviate the impact of allelic dropout on estimates of genoinformativity, we introduce a Zero-and-N-inflated Beta Binomial distribution parameterized by an additional variable ζ_j which defines the probability of allelic dropout for the gene.

$$P(C_{ij}^M, C_{ij}^P|H_{ij},S_j,G_j,\beta_j,\zeta_j) = \begin{cases} \frac{\zeta_j\alpha_j}{\alpha_j+\beta_j} + (1-\zeta_j)\text{Beta-Binomial}(\alpha_j,\beta_j,D_{ij}) & C_{ij}^M = 0 \\ \frac{\zeta_j\beta_j}{\alpha_j+\beta_j} + (1-\zeta_j)\text{Beta-Binomial}(\alpha_j,\beta_j,D_{ij}) & C_{ij}^P = 0 \\ \text{Beta-Binomial}(\alpha_j,\beta_j,D_{ij}) & \text{otherwise} \end{cases} \tag{17}$$

6.7.1 Implementation

Haplotype Inference Unfortunately due to inherent noise in the system and the cost of sampling the aforementioned Markov chain, we do not compute the Markov chain for each gene independently. Instead, we bin the genes into buckets B and perform a similar inference task with each bucket k. Each bucket on average contained 10 genes in our fits.

$$P(H_{ik}|\theta) = P(H_{ik}|H_{ik-1})P(H_{ik-1}|\theta) \tag{18}$$

We also used a fixed recombination rate r for each cell and each chromosome with the assumption that a cell would have on average 0.5 recombination events per chromosome.

Genoinformativity Inference Instead of learning the parameter S_j for each gene, we use a empirical estimate of S_j derived from dividing the number of H_M reads for a gene j by the total number of reads for that gene across all cells. We also tested using the mean of the empirical S_{ij} derived from each cell separately, and did not notice large differences in the model fits.

Priors

$$P(H_{ij}) = \begin{cases} 0.5 & H_{ij} = H_M \\ 0.5 & H_{ij} = H_P \end{cases} \tag{19}$$

$$P(G_j) = \text{Uniform}(0, 1) \tag{20}$$

$$P(\beta_j) = \text{Uniform}(3, 30) \tag{21}$$

$$P(\zeta_j) = \text{Uniform}(0.005, 1) \tag{22}$$

Two-stage Fitting For computational efficiency, we split the inference task into two stages. In the first stage, we fit both the haplotype and genoinformativity inference steps for highly expressed genes (TPM > 20). Then, in the second stage, we only performed genoinformativity inference using fixed haplotype probabilities. We used the mean posterior of the haplotypes from the first stage, and interpolated the probability for genes that were unique to the second stage. There was 99% correlation between the posterior mean of the genoinformativity values, indicating low variance in the posterior haplotype distributions and high confidence in haplotype inference.

Samplers We used PyMC3 (Salvatier, Wiecki, and Fonnesbeck 2016) as the framework for sampling the model. For the haplotype sampling, we used a Categorical Gibbs

Metropolis sampler. All the other parameters were sampled using the No U-turn Sampler (NUTS) with a target accept probability of 0.8. We sampled the model for 5000 steps with two separate chains and used the last 500 steps for estimating the posterior distribution across the 2 chains.

6.7.2 Sex chromosome GIMs

Mouse gene-level transcripts per million (TPM) values were collected for all genes in all spermatids using all RNAseq reads, not only allele-informative reads. For each gene, a loess regression was used to fit its log2 expression across the diffusion pseudotime with a pseudocount of 1 TPM, using the R loess function with a gaussian function family and 0.75 span. The residuals from this fit were then used to calculate pairwise Spearman correlations between all sex chromosome genes. Pairwise correlations were hierarchically clustered using the complete linkage method, with the results visualized in heatmaps. A cutoff height of 6 was empirically found to split the data into three clusters: a distinct X cluster, an anti-correlated distinct Y cluster, and a mixed X and Y cluster with no strong correlation patterns. Genes in the first two clusters were considered potential GIMs. We calculated the mean pairwise Spearman correlation between pairs of potential GIMs, with the sign reversed for genes in opposite clusters. Genes with a mean pairwise correlation of greater than 0.05 (roughly the median value over potential GIMs) were selected as putative sex chromosome GIMs.

6.8 GIM classification

To classify each gene as a "Confident GIM", "Confident Non-GIM", or "Remaining Gene", we fit the Bayesian model to shuffled data, and compared the posterior distributions for H_ik, G_j, and β_j between real and shuffled data. We utilized two main shuffling methods: complete shuffle and cell-label shuffle for each chromosome independently. The complete shuffle shuffled the allele counts randomly across the population of cells and genes. For the

cell-label shuffle, the allele counts were randomized across the cells, but the distribution of counts in a gene remained the same. We trained our Bayesian HMM using the same default parameters and priors as the real data, and compared the model fits. Since β_j can capture both the variance of single cell rna-seq as well as the variance in genoinformativity, we created a new measure γ_j as an alternative measure of genoinformativity that combines both posterior mean estimates of G_j and β_j.

$$\alpha_j = \frac{G_j(\beta_j)}{1 - G_j}$$
$$\gamma_j = \text{CDF of Beta}(\alpha_j, \beta_j) == 0.05$$

To reflect the confidence of the haplotype fits H_{ik} across all n samples, we also created an aggregated measure, fraction of poor haplotypes $\zeta_j = \frac{\sum_{i=1}^{N}\sum_{k=1}^{K} I[(H_{ik}<0.95)|H_ik>0.05)}{NK}$, which reflected the proportion of haplotypes that a posterior mean haplotype probability less than 0.95 for either the maternal or paternal haplotypes.

We performed a grid search across thresholds for highest posterior density (hpd) evaluated at 5% and 95% for genoinformativity G_j and $\gamma_j$andζ_j, which controlled the eFDR at 10% for confident and non-confident gims compared to the shuffled control. For a particular gene, the thresholds for a "Confident GIM" are: hpd 5% of genoinformativity > 0.025, hpd 95% of genoinformativity > 0.2, $\gamma_j > 0.025$, fraction of poor haplotypes < 0.4 For "Confident Non-GIMs" are restricted to hpd 95% of genoinformativity < 0.2 . Genes that fall outside these bounds were considered "Remaining Genes".

6.9 GIM characterization

Expression-matched control selection The expression trajectory across spermiogenesis was first tabulated for each gene by cross-referencing the log2 of the TPM expression level (with a pseudocount of 1 and complete dropout considered a zero) against the diffu-

sion map pseudotime value for each cell (i.e. the first dimension of the map). To reduce noise at the single cell level, a smoothed loess fit was used as the expression trajectory (fit using default parameters for the R `loess` function).

Next, all confident non-GIMs expressed in spermatids were considered as controls for all GIMs. Pools of controls were first reduced for each GIM based on two hard filters: first, all genes were equally distributed into 5 bins based on their dropout rates; second, the slope of a linear fit to the expression trajectory was required to differ by no more than 0.2. This helped to control for any confounders resulting in oversampling, as well as large expression changes in a small number of cells, generally in the extreme early or late part of the trajectory.

For each GIM, all non-GIMs remaining in its pool were ranked by their mean squared difference in log2 expression level, and the top 20 were selected as mock GIM controls, whose ranks were then scrambled. This resulted in 20 control sets of mock GIMs having similar dropout rates, slope of expression trajectory, and low difference in expression trajectory. For analyses limited to protein coding genes, control selection was performed again with both the GIMs and the control pools limited to protein coding genes.

For the cynomolgus samples, the expression trajectories were averaged across the two individuals. Where stringtie genes overlapped with Ensembl annotations, the aggregated expression for the Ensembl annotation was used for both GIMs and controls. A gene was considered a GIM if it was called as a confident GIM in either individual, and was considered as a non-GIM if it was called as a confident non-GIM in either indivdual. The rare genes having conflicting calls in each individual were excluded from these analyses. Human GIMs and non-GIMs were inferred from homologous cynomolgus annotations with homology defined as having the same Ensembl gene symbol (i.e. standard gene name).

For spermatid-expressed non-GIM controls, control sets were selected from all confident non-GIMs randomly, without filtering for dropout bin or expression trajectory fit.

Gene Ontology Mouse gene ontology annotations were downloaded from Ensembl Biomart with the Ensembl Genes 93 / GRCm38.p6 annotation dataset. The mean and standard deviation of number of GIMs expected with each annotation was calculated based on the 20 control sets. Nominal probabilities were then calculated using the normal distribution, and multiple testing was corrected using the Benjamini-Hochberg method to result in false discovery rates. GO terms were considered significant if they had at least 20 GIMs, an FDR ≤ 0.001 and a moderated log2 enrichment (using a pseudocount of 5) of at least 0.5.

For COMPARTMENTS comparisons, fewer controls had at least one annotation than GIMs, which could artificially inflate significance for individual categories. Therefore, we performed an additional normalization for the expected number of GIMs with an annotation. The number of controls in a set having a GO annotation was converted to a fraction out of those have any annotation, and then multiplied by the number of GIMs to yield the total number expected with each annotation specifically. Otherwise the enrichment analysis was the same as for the GO analysis above.

3′ UTR motifs Only protein-coding genes were considered. The 3′ UTR annotations of GIMs and their controls were taken from the highest expressed Ensembl transcript in spermatids. UTRs annotated as less than 7 nucleotides in length were discarded. All 20 sets of control UTRs were combined into a single background set, allowing duplicates. AME, a tool from the MEME suite, was run with default parameters using a motif database comprised of the CISBP-RNA and Ray2013 mouse and human sets provided by MEME. The enrichment search was performed using GIMs as foreground and the combined control set as background, with foreground and background switched for the depletion analysis.

For candidate RNA-binding proteins, only those with a maximum TPM of 10 at any point in the loess-smoothed expression trajectory were considered. Enrichments were considered significant at an E-value cutoff of 0.01. Motifs having the same IUPAC consensus

were merged into a single result.

Selective sweeps Candidates for mouse selective sweeps were taken from Staubach et al. 2012. Sweep regions in any population were considered. All candidate genes within 600kb of each other were collapsed into a single region. For GIMs or each control set, the number of regions having at least one overlapping gene was counted. For a p-value of this difference, the mean and standard deviation of the control sets was used to generate a one-sample t-test.

Candidates for human selective sweeps were taken from Refoyo-Martínez et al. 2019; Schrider and Kern 2016; Ferrer-Admetlla et al. 2014; Cheng, Racimo, and Nielsen 2019; Munch et al. 2016, with selective sweep regions as in each paper. In cases in which the paper predicted selective sweep regions but did not annotate associated genes, all genes overlapping the regions were considered selective sweep candidates. Otherwise this analysis was as in the mouse.

Direct testing for human selective sweeps was performed using statistics from Pybus et al. 2014 based on analysis of the 1000 genomes project data. To help control for differences in gene length, the median score overlapping the $3'$ UTR was used to represent the gene. The "best" score for each gene was taken across each population, where "best" signifies the raw score most in favor of a selective sweep for that score. Selective sweep candidates were defined as any where the score was at least 3 standard deviations beyond the mean in this direction. The number of GIM sweep candidates was compared to background expectation of the mean and standard deviation among the 20 control sets.

Testis-specific paralogs Paralog and tissue-specificity data were taken from Guschanski, Warnefors, and Kaessmann 2017. Testis-specific paralogs were defined as those with a "Tissue specificity" (as defined by the paper) of at least 0.90.

Alternative splicing Mouse alternative splicing was taken from events in VastDB with quality greater than zero and testis specific was defined as a difference in PSI of at least 50 between testis and the median PSI across all other tissues.

Human isoform expression was taken from the GTEX consortium (file `GTEx_Analysis_2016-01-15_v7_`
Since individual isoform estimates can be unstable, we considered subsets of isoforms that are expressed higher in testis. Each transcript was ranked by the difference between testis isoform usage (i.e. ratio of transcript TPM to gene TPM in that tissue) to the median tissue isoform usage across other tissues. The maximum of the cumulative sum of excess isoform usage in testis was counted as the testis specificity (testis isoform usage minus other tissue isoform usage). A cutoff of 0.5 was considered testis-specific (equivalent to 50 PSI).

Late translation Translation data was taken from Iguchi, Tobias, and Hecht 2006 (GSE4711 on GEO). Translation efficiencies were calculated as the median across replicates of the fold change from polysome to RNP samples. Genes were defined as having specific late translation if they were in the bottom quartile of this score at day 22 (which is depleted for late spermiogenesis), and the top quartile with respect to fold-change increase in translation efficiency between day 22 and adult mice. For each of the functional readouts of GIMs (e.g. selective sweeps), we compared the fraction of GIMs in that category that were specifically late translated to those that were not in that category (i.e. not functional candidates by that measure).

References

Alavioon2017 Ghazal Alavioon et al. "Haploid selection within a single ejaculate increases offspring fitness." In: *Proceedings of the National Academy of Sciences of the United States of America* 114.30 (July 2017), pp. 8053–8058. ISSN: 1091-6490. DOI: 10.1073/pnas.1705601114. URL: http://www.ncbi.nlm.nih.gov/

pubmed/28698378%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5544320.

Andreassi2009 Catia Andreassi and Antonella Riccio. "To localize or not to localize: mRNA fate is in 3UTR ends". In: *Trends in Cell Biology* 19.9 (Sept. 2009), pp. 465–474. ISSN: 09628924. DOI: 10.1016/j.tcb.2009.06.001. URL: http://www.ncbi.nlm.nih.gov/pubmed/19716303%20https://linkinghub.elsevier.com/retrieve/pii/S096289240900141X.

Angerer2016 Philipp Angerer et al. "destiny : diffusion maps for large-scale single-cell data in R". In: *Bioinformatics* 32.8 (Apr. 2016), pp. 1241–1243. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv715. URL: http://www.ncbi.nlm.nih.gov/pubmed/26668002%20https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv715.

Bastos2005 Henri Bastos et al. "Flow cytometric characterization of viable meiotic and postmeiotic cells by Hoechst 33342 in mouse spermatogenesis". In: *Cytometry Part A* 65A.1 (May 2005), pp. 40–49. ISSN: 1552-4922. DOI: 10.1002/cyto.a.20129. URL: http://doi.wiley.com/10.1002/cyto.a.20129.

Binder2014 J. X. Binder et al. "COMPARTMENTS: unification and visualization of protein subcellular localization evidence". In: *Database* 2014.0 (Feb. 2014), bau012–bau012. ISSN: 1758-0463. DOI: 10.1093/database/bau012. URL: https://academic.oup.com/database/article-lookup/doi/10.1093/database/bau012.

Braun1989 Robert E. Braun et al. "Genetically haploid spermatids are phenotypically diploid". In: *Nature* 337.6205 (Jan. 1989), pp. 373–376. ISSN: 00280836. DOI: 10.1038/337373a0. URL: http://www.nature.com/articles/337373a0.

Butler2007 Avigdor Butler et al. "Sperm abnormalities in heterozygous acid sphingomyelinase knockout mice reveal a novel approach for the prevention of genetic diseases." In: *The American journal of pathology* 170.6 (June 2007), pp. 2077–88.

947 ISSN: 0002-9440. DOI: 10.2353/ajpath.2007.061002. URL: http://www.ncbi.
948 nlm.nih.gov/pubmed/17525274%20http://www.pubmedcentral.nih.gov/
949 articlerender.fcgi?artid=PMC1899442.

950 **Buxbaum2015** Adina R Buxbaum, Gal Haimovich, and Robert H Singer. "In the right
951 place at the right time: visualizing and understanding mRNA localization." In:
952 *Nature reviews. Molecular cell biology* 16.2 (Feb. 2015), pp. 95–109. ISSN: 1471-
953 0080. DOI: 10.1038/nrm3918. URL: http://www.ncbi.nlm.nih.gov/pubmed/
954 25549890%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?
955 artid=PMC4484810.

956 **Cheng2019** Jade Yu Cheng, Fernando Racimo, and Rasmus Nielsen. "Ohana: detecting
957 selection in multiple populations by modelling ancestral admixture components".
958 In: *bioRxiv* (Feb. 2019), p. 546408. DOI: 10.1101/546408. URL: https://www.
959 biorxiv.org/content/10.1101/546408v1.

960 **Cingolani2012** Pablo Cingolani et al. "A program for annotating and predicting the
961 effects of single nucleotide polymorphisms, SnpEff". In: *Fly* 6.2 (Apr. 2012), pp. 80–
962 92. ISSN: 1933-6934. DOI: 10.4161/fly.19695. URL: http://www.tandfonline.
963 com/doi/abs/10.4161/fly.19695.

964 **Cocquet2012** Julie Cocquet et al. "A Genetic Basis for a Postmeiotic X Versus Y Chro-
965 mosome Intragenomic Conflict in the Mouse". In: *PLoS Genetics* 8.9 (Sept. 2012).
966 Ed. by Michael W. Nachman, e1002900. ISSN: 1553-7404. DOI: 10.1371/journal.
967 pgen.1002900. URL: https://dx.plos.org/10.1371/journal.pgen.1002900.

968 **Conway1994** S. J. Conway et al. "Y353/B: a candidate multiple-copy spermiogene-
969 sis gene on the mouse Y chromosome". In: *Mammalian Genome* 5.4 (Apr. 1994),
970 pp. 203–210. ISSN: 0938-8990. DOI: 10.1007/BF00360546. URL: http://link.
971 springer.com/10.1007/BF00360546.

Cox2009 Allison Cox et al. "A new standard genetic map for the laboratory mouse". In: *Genetics* 182.4 (Aug. 2009), pp. 1335–1344. ISSN: 00166731. DOI: 10.1534/genetics.109.105486. URL: http://www.ncbi.nlm.nih.gov/pubmed/19535546%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2728870.

Crow1979 J F Crow. "Genes That Violate Mendel Rules". In: *Scientific American* 240.2 (1979), pp. 134–&. DOI: 10.2307/24965130. URL: https://www.jstor.org/stable/24965130.

Eep2019 Johnson Eep, Ellis Pji, and Peter Ellis Email. "Differential sperm motility mediates the sex ratio drive shaping mouse sex chromosome evolution". In: (). DOI: 10.1101/649707. URL: http://dx.doi.org/10.1101/649707.

Ferrer2014 Anna Ferrer-Admetlla et al. "On detecting incomplete soft or hard selective sweeps using haplotype structure." In: *Molecular biology and evolution* 31.5 (May 2014), pp. 1275–91. ISSN: 1537-1719. DOI: 10.1093/molbev/msu077. URL: http://www.ncbi.nlm.nih.gov/pubmed/24554778%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3995338.

Gaysinskaya2014 Valeriya Gaysinskaya et al. "Optimized flow cytometry isolation of murine spermatocytes". In: *Cytometry Part A* 85.6 (June 2014), pp. 556–565. ISSN: 15524922. DOI: 10.1002/cyto.a.22463. URL: http://doi.wiley.com/10.1002/cyto.a.22463.

Guschanski2017 Katerina Guschanski, Maria Warnefors, and Henrik Kaessmann. "The evolution of duplicate gene expression in mammalian organs". In: *Genome Research* 27.9 (Sept. 2017), pp. 1461–1474. ISSN: 1088-9051. DOI: 10.1101/gr.215566.116. URL: http://www.ncbi.nlm.nih.gov/pubmed/28743766%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5580707%20http://genome.cshlp.org/lookup/doi/10.1101/gr.215566.116.

Iguchi2006 Naoko Iguchi, John W Tobias, and Norman B Hecht. "Expression profiling reveals meiotic male germ cell mRNAs that are translationally up- and down-regulated." In: *Proceedings of the National Academy of Sciences of the United States of America* 103.20 (May 2006), pp. 7712–7. ISSN: 0027-8424. DOI: 10.1073/pnas.0510999103. URL: http://www.ncbi.nlm.nih.gov/pubmed/16682651%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1472510.

Immler2008 Simone Immler. "Sperm competition and sperm cooperation: the potential role of diploid and haploid expression." In: *Reproduction (Cambridge, England)* 135.3 (Mar. 2008), pp. 275–83. ISSN: 1741-7899. DOI: 10.1530/REP-07-0482. URL: http://www.ncbi.nlm.nih.gov/pubmed/18299420.

Joseph2004 Sarah B. Joseph and Mark Kirkpatrick. *Haploid selection in animals*. Nov. 2004. DOI: 10.1016/j.tree.2004.08.004. URL: https://www.sciencedirect.com/science/article/pii/S0169534704002381.

Keane2011 Thomas M. Keane et al. "Mouse genomic variation and its effect on phenotypes and gene regulation". In: *Nature* 477.7364 (Sept. 2011), pp. 289–294. ISSN: 0028-0836. DOI: 10.1038/nature10413. URL: http://www.nature.com/articles/nature10413.

Kleene2005 Kenneth C. Kleene. "Sexual selection, genetic conflict, selfish genes, and the atypical patterns of gene expression in spermatogenic cells". In: *Developmental Biology* 277.1 (Jan. 2005), pp. 16–26. ISSN: 0012-1606. DOI: 10.1016/J.YDBIO.2004.09.031. URL: https://www.sciencedirect.com/science/article/pii/S0012160604006682.

Lecuyer2007 Eric Lécuyer et al. "Global Analysis of mRNA Localization Reveals a Prominent Role in Organizing Cellular Architecture and Function". In: *Cell* 131.1 (Oct. 2007), pp. 174–187. ISSN: 00928674. DOI: 10.1016/j.cell.2007.08.003. URL: http://www.ncbi.nlm.nih.gov/pubmed/17923096%20http://linkinghub.elsevier.com/retrieve/pii/S0092867407010227.

Li2009 H. Li and R. Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform". In: *Bioinformatics* 25.14 (July 2009), pp. 1754–1760. DOI: 10.1093/bioinformatics/btp324. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp324.

Li2009a H. Li, B. Handsaker, et al. "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16 (Aug. 2009), pp. 2078–2079. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp352. URL: http://www.ncbi.nlm.nih.gov/pubmed/19505943%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2723002%20https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352.

Liao2014 Y. Liao, G. K. Smyth, and W. Shi. "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features". In: *Bioinformatics* 30.7 (Apr. 2014), pp. 923–930. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt656. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt656.

Liu2014 Eric Yi Liu et al. "High-Resolution Sex-Specific Linkage Maps of the Mouse Reveal Polarized Distribution of Crossovers in Male Germline". In: *Genetics* 197.1 (May 2014), pp. 91–106. ISSN: 0016-6731. DOI: 10.1534/GENETICS.114.161653. URL: https://www.genetics.org/content/197/1/91.short.

Martin-DeLeon2005 Patricia A. Martin-DeLeon et al. "Spam I-associated transmission ratio distortion in mice: Elucidating the mechanism". In: *Reproductive Biology and Endocrinology* 3.1 (2005), p. 32. ISSN: 14777827. DOI: 10.1186/1477-7827-3-32. URL: http://rbej.biomedcentral.com/articles/10.1186/1477-7827-3-32.

Meikar2014 Oliver Meikar et al. "An atlas of chromatoid body components." In: *RNA (New York, N.Y.)* 20.4 (2014), pp. 483–95. ISSN: 1469-9001. DOI: 10.1261/rna.043729.113. URL: http://www.ncbi.nlm.nih.gov/pubmed/24554440.

Morales2002 Carlos R. Morales et al. "A TB-RBP and Ter ATPase Complex Accompanies Specific mRNAs from Nuclei through the Nuclear Pores and into Intercellular Bridges in Mouse Male Germ Cells". In: *Developmental Biology* 246.2 (June 2002), pp. 480–494. ISSN: 0012-1606. DOI: 10.1006/DBIO.2002.0679. URL: https://www.sciencedirect.com/science/article/pii/S0012160602906792.

Munch2016 Kasper Munch et al. "Selective Sweeps across Twenty Millions Years of Primate Evolution". In: *Molecular Biology and Evolution* 33.12 (Dec. 2016), pp. 3065–3074. ISSN: 0737-4038. DOI: 10.1093/molbev/msw199. URL: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msw199.

Nadeau2017 Joseph H Nadeau. "Do Gametes Woo? Evidence for Their Nonrandom Union at Fertilization." In: *Genetics* 207.2 (Oct. 2017), pp. 369–387. ISSN: 1943-2631. DOI: 10.1534/genetics.117.300109. URL: http://www.ncbi.nlm.nih.gov/pubmed/28978771.

Pybus2014 Marc Pybus et al. "1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans". In: *Nucleic Acids Research* 42.D1 (Jan. 2014), pp. D903–D909. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1188. URL: http://www.ncbi.nlm.nih.gov/pubmed/24275494%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3965045%20https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1188.

Refoyo-Martinez2019 Alba Refoyo-Martínez et al. "Identifying loci under positive selection in complex population histories." In: *Genome research* (July 2019), gr.246777.118. ISSN: 1549-5469. DOI: 10.1101/gr.246777.118. URL: http://www.ncbi.nlm.nih.gov/pubmed/31362936.

Ryder2018 Pearl V. Ryder and Dorothy A. Lerit. "RNA localization regulates diverse and dynamic cellular processes". In: *Traffic* 19.7 (July 2018), pp. 496–502. ISSN:

13989219. DOI: 10.1111/tra.12571. URL: http://doi.wiley.com/10.1111/tra.12571.

Salvatier2016 John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. "Probabilistic programming in Python using PyMC3". In: *PeerJ Computer Science* 2 (Apr. 2016), e55. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.55. URL: https://peerj.com/articles/cs-55.

Schrider2016 Daniel R. Schrider and Andrew D. Kern. "S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning". In: *PLOS Genetics* 12.3 (Mar. 2016). Ed. by Bret A. Payseur, e1005928. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1005928. URL: https://dx.plos.org/10.1371/journal.pgen.1005928.

Staubach2012 Fabian Staubach et al. "Genome Patterns of Selection and Introgression of Haplotypes in Natural Populations of the House Mouse (Mus musculus)". In: *PLoS Genetics* 8.8 (Aug. 2012). Ed. by Michael H. Kohn, e1002891. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1002891. URL: http://dx.plos.org/10.1371/journal.pgen.1002891.

Tarasov2015 Artem Tarasov et al. "Sambamba: fast processing of NGS alignment formats". In: *Bioinformatics* 31.12 (June 2015), pp. 2032–2034. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv098. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv098.

Umehara2019 Takashi Umehara, Natsumi Tsujita, and Masayuki Shimada. "Activation of Toll-like receptor 7/8 encoded by the X chromosome alters sperm motility and provides a novel simple technology for sexing sperm". In: *PLOS Biology* 17.8 (Aug. 2019). Ed. by Yukiko M Yamashita, e3000398. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.3000398. URL: http://dx.plos.org/10.1371/journal.pbio.3000398.

Ventela2003 Sami Ventelä, Jorma Toppari, and Martti Parvinen. "Intercellular organelle traffic through cytoplasmic bridges in early spermatids of the rat: mechanisms of haploid gene product sharing." In: *Molecular biology of the cell* 14.7 (July 2003), pp. 2768–80. ISSN: 1059-1524. DOI: 10.1091/mbc.E02-10-0647. URL: http://www.ncbi.nlm.nih.gov/pubmed/12857863%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC165675.

Veron2009 Nathalie Véron et al. "Retention of gene products in syncytial spermatids promotes non-Mendelian inheritance as revealed by the t complex responder." In: *Genes & development* 23.23 (Dec. 2009), pp. 2705–10. ISSN: 1549-5477. DOI: 10.1101/gad.553009. URL: http://www.ncbi.nlm.nih.gov/pubmed/19952105%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2788329.

Zheng2001 Y. Zheng, X. Deng, and P.A. Martin-DeLeon. "Lack of Sharing of Spam1 (Ph-20) among Mouse Spermatids and Transmission Ratio Distortion1". In: *Biology of Reproduction* 64.6 (June 2001), pp. 1730–1738. ISSN: 0006-3363. DOI: 10.1095/biolreprod64.6.1730. URL: https://academic.oup.com/biolreprod/article-lookup/doi/10.1095/biolreprod64.6.1730.

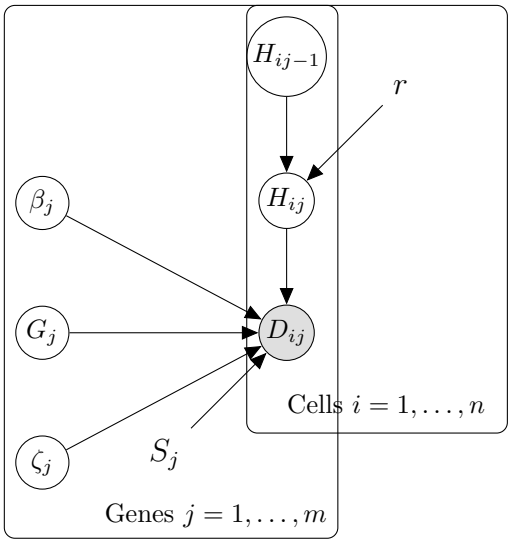


Figure 5: Graphical model for Bayesian method