Mutational bias in spermatogonia impacts the anatomy of regulatory sites in the human genome

Vera B. Kaiser[1], Lana Talmane[1], Yatendra Kumar[1], Fiona Semple[1], Marie MacLennan[1], Deciphering Developmental Disorders Study[1,2], David R. FitzPatrick[1], Martin S. Taylor[1*], Colin A. Semple[1*]

[1]MRC Human Genetics Unit, MRC Institute of Genetics and Cancer, The University of Edinburgh, Western General Hospital, Crewe Road South, Edinburgh EH4 2XU, UK

[2]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

*Equal contribution

Corresponding author: Vera B Kaiser vera.kaiser@ed.ac.uk

Keywords: germline structural variation, ATAC-seq, regulatory genomics, spermatogonia, PRDM9, NRF1

Abstract

Mutation in the germline is the ultimate source of genetic variation, but little is known about the influence of germline chromatin structure on mutational processes. Using ATAC-seq, we profile the open chromatin landscape of human spermatogonia, the most proliferative cell-type of the germline, identifying transcription factor binding sites (TFBSs) and PRDM9-binding sites, a subset of which will initiate meiotic recombination. We observe an increase in rare structural variant (SV) breakpoints at PRDM9-bound sites, implicating meiotic recombination in the generation of structural variation. Many germline TFBSs, such as NRF, are also associated with increased rates of SV breakpoints, apparently independent of recombination. Singleton short insertions (>=5 bp) are highly enriched at TFBSs, particularly at sites bound by testis active TFs, and their rates correlate with those of structural variant breakpoints. Short insertions often duplicate the TFBS motif, leading to clustering of motif sites near regulatory regions in this male-driven evolutionary process. Increased mutation loads at germline TFBSs disproportionately affect neural enhancers with activity in spermatogonia, potentially altering neurodevelopmental regulatory architecture. Local chromatin structure in spermatogonia is thus pervasive in shaping both evolution and disease.

Introduction

Mutation is the ultimate source of genetic variation, and inherited variation must invariably arise in the germline. It is well established from cross-species comparisons that the rate of nucleotide substitution mutations fluctuates at the multi-megabase

(>10^6 bp) scale across the genome (Wolfe et al. 1989; Hodgkinson and Eyre-Walker 2011), with early replicating regions subject to reduced rates of mutation. These patterns similarly manifest in the rate of human single nucleotide polymorphisms (SNPs) (Stamatoyannopoulos et al. 2009). Germline structural variation in the human genome is also associated with replication timing, such that copy number variants (CNVs) emerging from homologous recombination-based mechanisms are enriched in early replicating regions, while CNVs arising from non-homologous mechanisms are enriched in late replicating regions (Koren et al. 2012). Local chromatin structure also influences the mutation rate. However, finer-scale variation (<1Mb) in the germline mutation rate has so far only been related to genomic features derived from somatic cells (Gonzalez-Perez et al. 2019) because human germline-derived measures of chromatin structure have only recently become available (Guo et al. 2017; Guo et al. 2018). Transcription factor binding sites (TFBSs) are particularly prone to point mutations in cancer (Kaiser et al. 2016), probably due to interference between TF binding and the replication and repair machinery (Reijns et al. 2015; Sabarinathan et al. 2016; Afek et al. 2020), but the mutational consequences of binding at these sites in the germline is unknown.

During meiosis, homologous recombination may introduce short mutations or render genomic regions prone to rearrangements (Pratto et al. 2014; Halldorsson et al. 2019). A key player in this process is PRDM9, which binds its cognate sequence motif and directs double-strand break (DSB) formation in meiotic prophase (Baudat et al. 2010; Myers et al. 2010). In humans, PRDM9 binding site occupancy has only been directly assayed in a somatic cell line (Altemose et al. 2017), whereas indirect measures of PRDM9 activity include a proxy for DSBs (DMC1-bound single stranded DNA

(ssDNA)) in testis (Pratto et al. 2014), and population genetic based measures of recombination hotspots (HSs) (Myers et al. 2005; The 1000 Genomes Project Consortium et al. 2015). The method ATAC-Seq (Buenrostro et al. 2013) reports chromatin accessibility and provides a snapshot of all active regulatory regions and occupied binding sites in a given tissue. In particular, ATAC-Seq footprinting (Sherwood et al. 2014; Li et al. 2019), when applied to spermatogonia, has the potential to reveal the binding of hundreds of TFs, as well as PRDM9, in the male germline. In addition, large human genome sequencing projects can be used to reveal patterns of mutation rates, by focussing on extremely rare variants (Messer 2009; Carlson et al. 2018; Li and Luscombe 2020). Making use of such variant datasets as well as novel ATAC-Seq data in spermatogonia, we study the mutational landscape at transcription factor binding sites (TFBSs) in accessible human spermatogonial chromatin.

Results

Spermatogonial regulatory regions are enriched for rare deletion breakpoints

ATAC-Seq (Buenrostro et al. 2013) reports local chromatin accessibility and provides a snapshot of active regulatory regions and genomic regions occupied by DNA-binding proteins in a given tissue. We used ATAC-Seq to identify open chromatin sites in FGFR3-positive spermatogonial cells isolated from dissociated human testicular samples. FGFR3 is most highly expressed in self-renewing spermatogonial stem cells, with low expression also being detected in early differentiating spermatogonia (Guo et al. 2018; Sohni et al. 2019); its expression thus overlaps with

the onset of PRDM9 expression in pre-meiotic spermatogonia (Human Protein Atlas: https://www.proteinatlas.org/ENSG00000164256-PRDM9/celltype/testis and https://www.proteinatlas.org/ENSG00000068078-FGFR3/celltype/testis) (Guo et al. 2018). Open chromatin in FGFR3-positive cells was identified using standard peak detection analysis (Methods) and multiple metrics (Supplemental Fig. S1) indicated high data quality (Yan et al. 2020). Hierarchical clustering (Ramirez et al. 2016) showed that this novel spermatogonial ATAC-Seq dataset displays a genome-wide distribution of peaks consistent with other spermatogonial derived data, and is distinct from ES cell and somatic tissue datasets (Supplemental Fig. S2).

Next, we assessed the enrichments of different classes of sequence variants at spermatogonial active sites, including singleton SV breakpoint frequencies as a proxy for the mutation rate of such variants. We made use of ultra-rare genomic variants from a variety of human sequencing studies: the Deciphering Developmental Disorders (DDD) study (Deciphering Developmental Disorders Study 2015; Mcrae et al. 2017) of severe and undiagnosed developmental disorders (https://www.ddduk.org/), a large collection of variants from an aggregated database (gnomAD; http://gnomad.broadinstitute.org/), and *de novo* variants from trio sequencing studies (http://denovo-db.gs.washington.edu/, https://research.mss.ng/, An et al. (2018)). Based on the DDD dataset - a combination of high-density arrayCGH and exome sequencing (Deciphering Developmental Disorders Study 2015) - we identified 6,704 singleton deletion variants among 9,625 DDD probands (carrier frequency of ~ 0.002) (Supplemental Table S1). Permutation analysis demonstrates that DDD singleton breakpoints are enriched at spermatogonial ATAC-Seq sites, their overlap being > 4-times the expected genome-wide rate (Supplemental Table S2), and

shifted permutation Z-scores reveal that the enrichment is specific to the ATAC-Seq peaks as opposed to wider genomic regions (Figure 1). We also considered 6,013 deletions (with 7,365 unique breakpoints) that were present in the DDD consensus dataset (Deciphering Developmental Disorders Study 2015) (Methods) at a frequency of at least 1%, representing variants expected to be relatively common in human populations (Supplemental Table S1). These variants show a dip in frequency and downward trend near active sites (Figure 1a). However, we note that the overlap between common variant breakpoints and ATAC-Seq peaks is still ~ 2-fold higher than the expected genome-wide rate ($p < 10^{-4}$). We conclude that singleton deletion breakpoints often occur at TFBSs in spermatogonia, suggesting a higher mutational input or less accurate repair at these sites compared to neighbouring regions. The breakpoints of more common variants are observed less frequently at the same binding sites, which may indicate the action of purifying selection in the removal of deleterious mutations at these active regulatory sites.

Similar trends are also observed for singleton deletion breakpoints from an independent large-scale aggregated dataset of human variants (Figure 1e, 1f) from whole genome sequence (WGS) analysis (Collins et al. 2020) (Supplemental Table S1). We again find a significant enrichment of singleton variant breakpoints at ATAC-Seq peaks, and this enrichment is not seen for common variants (Figure 1).

Locally elevated mutation at spermatogonial TFBSs

Compared to larger structural variants, such as those (up to megabase sized) deletions examined above, indels have been shown to occur at a higher rate of about 6 new variants per genome and generation (Besenbacher et al. 2016). Short indels (<= 4 bp)

are thought to arise due to replication slippage (Levinson and Gutman 1987; Montgomery et al. 2013), whereas longer variants have been considered a hallmark of inaccurate DNA repair after DSBs (Rodgers and McVey 2016). Here, we focus on gnomAD singleton indels <= 20 bp as these variants are expected to be well resolved using short read sequencing. To enable higher spatial resolution of the mutation patterns at ATAC-seq defined accessible chromatin regions, and for the subsequent inference of the associated DNA binding proteins, we identified 706,008 protein binding sites using ATAC-Seq footprinting analysis (Li et al. 2019) (Methods; Supplemental Tables S3 and S4). The rate of singleton 5-20 bp insertions at footprinted spermatogonial protein binding sites approximately doubles from background expectation and is highly concentrated to within 1 kb of the binding site (Figure 2); shifted Z-scores based on genome-wide circular permutations similarly show a highly localized spike of insertions around TFBSs (Figure 2). This pattern starkly contrasts the localised depletion of common variants of the same mutation class at the same binding sites (Figure 2), again implicating a locally elevated mutation rate and purifying selection. In fact, most classes of rare mutation (singleton SVs, smaller and longer indels, SNPs) are significantly enriched at spermatogonial TFBSs (Figure 3), and in the gnomAD dataset, where all singleton classes have been ascertained by WGS, the enrichment is strongest for insertions >= 5 bp. We confirmed the enrichment of singleton short insertions and SV deletion breakpoints at spermatogonial TFBSs, using an independent permutation approach with bedtools shuffle (Quinlan and Hall 2010) (Supplemental Table S5).

In addition to singleton variants from large population samples, we also compiled a set of “gold standard” *de novo* short variants from a range of trio sequencing studies (see Methods). The *de novo* variants show a strikingly similar

trend to the gnomAD singleton variants, with a moderate (~10-60%) increase of mutation rates at TFBSs for all categories of short 1-2bp sequence variants, a but larger increase of ~130% for insertions of 5-20 bp (Figure 3). These results were confirmed using a set of independent positive and negative control sites (Supplemental Fig. S3). We conclude that regulatory sites that are active in spermatogonia show unusual parallel enrichments for both large SV breakpoints and 5-20 bp insertions, consistent with localised DNA damage or error-prone repair.

Germline PRDM9 and NRF1 binding generate hotspots for structural variation

To examine any differences in mutational loads associated with different binding factors, we analysed mutational patterns stratified by the binding factors included in the JASPAR database (Sandelin et al. 2004). We accounted for redundancy caused by multiple factors binding to a single motif by considering 167 motif families (Supplemental Table S6). Furthermore, using the reported binding site motif for PRDM9 (Myers et al. 2008), we defined 9,778 putative PRDM9-bound sites corroborated by evidence for H3K4me3 enrichment in testis (Methods).

The spermatogonial binding sites of 11% (19/167) of motif families overlapped DDD singleton deletion breakpoints more often than expected (Bonferroni corrected p = 0.017) and no motif family was found to be depleted for breakpoints (Supplemental Table S3), suggesting that increased load is a common feature of TFBSs bound by different transcription factors in the germline. Similarly, singleton 5-20bp insertions from the gnomAD database were found to be significantly enriched at 29% (48/167) of families (Bonferroni corrected p = 0.017) and, nominally, 84% (140/167) of families showed enrichment for these insertions (Supplemental Table S4). Again, no

TFBS family was found to be depleted for these rare variants. Collectively, these results suggest that TFBSs active in spermatogonia incur locally elevated burdens of short insertions and large structural variants across many different binding motifs.

Certain motif families appear to carry notably higher mutational loads than the general disruption seen across all TFBSs. Based on the insertion fold enrichment (IFE), i.e. the ratio of the observed to expected numbers of insertions (5-20 bp), PRDM9 binding sites are among the most disrupted sites in the genome (IFE = 6.3), and this also holds for PRDM9 sites outside known sites of recombination (IFE = 6.7 for 8,139 PRDM9 sites with a distance of at least 500bp from HSs and ssDNA sites, respectively). PRDM9 sites are similarly associated with higher rates of singleton deletion breakpoints (Figure 4a, 4c), in line with the roles of PRDM9 during recombination, though PRDM9 sites outside known sites of recombination also show this trend (observed overlaps with deletion breakpoints = 9; expected = 1; $p < 10^{-4}$). Two other TFBS families, corresponding to NRF1 (Nuclear Respiratory Factor 1; IFE=7.0) and HINFP (IFE=6.6) exceed the disruption seen at PRDM9 sites, and remarkably, NRF1 sites are disrupted at high rates according to DDD breakpoint data (Supplemental Table S3). Shifted Z-scores for the enrichment of short insertions (5-20 bp) at both NRF1 and PRDM9 binding sites are in the top four, next to SP/KLF transcription factors (motif families 938 and 992), suggesting strong focal enrichments at these sites (Supplemental Tables S4 and S6). NRF1 has been shown to be an important testis-expressed gene with meiosis-specific functions (Wang et al. 2017; Palmer et al. 2019), but NRF1 binding sites have, to our knowledge, not been reported to be foci for genomic instability. We find strikingly similar enrichments of short insertions (5-20 bp) at TFBSs in SSEA4- and KIT-marked spermatogonial

samples produced in previous ATAC-seq studies (Guo et al. 2017; Guo et al. 2018). Reprocessing these previous datasets identically to our own reveals that PRDM9, NRF1 and HINFP sites are again among the top 5 disrupted motif families (Supplemental Tables S7 and S8).

Although both PRDM9 and NRF1 binding sites are GC-rich, their modest motif similarity suggests that the two factors occupy distinct binding motifs (PWMclus: Pearson's correlation distance $r = 0.35$ for PRDM9 *versus* and NRF1) and should not converge on the same sites. However, in practice, PRDM9 and NRF1 binding sites were often found within the same regulatory regions, such that many (1,199) ATAC-Seq peaks contained both the NRF1 and PRDM9 binding motifs. The disruption of motifs within these co-bound peaks was notably higher, with NRF1-motifs being disrupted by short insertions 10.8-fold the expected rate (observed: 108; expected: 10), and PRDM9-motifs 11.2-fold the expected rate (observed: 146; expected: 13) when co-occurring with the other factor ($p < 10^{-4}$ in each case). Similarly, 1,311 ATAC-Seq peaks contained a motif for both CTCF and PRDM9, and CTCF motifs in these peaks were more highly disrupted by short insertions (ratio = 6.3; observed: 69; expected: 11) compared to all CTCF motifs (Supplemental Table S4), as was PRDM9 (ratio = 8.2; observed: 115; expected: 14) ($p < 10^{-4}$ in each case).

Importantly, the excess of insertions observed at particular motif sites is not a trivial consequence of statistical power (i.e. the number of TFBSs in the genome); for example, the number of binding sites identified for PRDM9 and NRF1 is fewer than many other factors (< 10,000 sites each; Supplemental Tables S3 and S4).

In general, mutational loads appear to be dependent on the level of chromatin accessibility (MACS2 peak scores) and the number of factors predicted to bind at

ATAC-Seq defined regulatory regions, such that regions in the upper quartile of accessibility that are also occupied by more than 4 factors incur the highest indel loads (Supplemental Fig. S4). The significant positive correlation between the rates of binding site disruption via singleton insertions and deletion breakpoints across all motif families (Supplemental Fig. S5; Spearman's R = 0.52, $p < 10^{-5}$) suggests that the two types of damage may be mechanistically linked. In support of this idea, singleton short insertions (5-20 bp) and singleton SV deletion breakpoints overlap at the exact nucleotide position more often than expected (genome-wide Z-score = 26.31; $p < 10^{-4}$; see also Supplemental Fig. S6). This overlap is unlikely to be due to erroneous variant calling in the singleton dataset since we observe similar patterns for common variants of the same variant categories (genome-wide Z-score = 62.9, $p < 10^{-4}$).

Short insertions generate clustered binding sites within regulatory regions

Intriguingly, the 5-20 bp insertions observed at TFBSs frequently occur within only a few nucleotides of the binding motifs, whereas other classes of short variants do not show such a precisely localized increase (Figure 5 and Supplemental Fig. S7). Despite a moderate genome-wide enrichment (Figure 3), the 1-2 bp insertions characteristic of polymerase slippage, do not peak in the immediate neighbourhood of TFBSs (Figure 5 and Supplemental Fig. S7). We examined the consequences of elevated 5-20 bp insertion rates at TFBSs using an exhaustive motif search algorithm (Bailey et al. 2009), which finds overrepresented sequence motifs among a set of input sequences. We found that the inserted sequences at a mutated TFBS often contain additional copies of the sequence motif corresponding to the original TFBS (Figure 6 and

Supplemental Fig. S8), suggesting that many insertions at TFBSs are tandem duplication events, including events at CTCF, NRF1 and PRDM9 sites. The presence of these motif-containing singleton insertions appears to reveal a novel mutational mechanism expected to increase the number of binding sites for a binding factor and to lead to the expansion of TFBS clusters. CTCF-binding sites are known to occur in clusters (Kentepozidou et al. 2020) and are often affected by singleton insertions in our dataset (ranked 12th out of 167 motif families, based on the number of insertions per TFBS; Supplemental Table S4). We find that spermatognial active sites exhibit greater homotypic clustering of TFBS than ATAC-seq defined binding sites from somatic tissues (Figure 6). Combined with a positive correlation between homotypic motif clustering and insertion rate, this suggests that spermatogonia binding sites are progressively accruing motif clusters.

These unusual patterns of clustered TFBSs at indel breakpoints appear to be specific to spermatogonial ATAC-seq peaks, and do not reflect genome-wide trends. Applying the MEME-Chip algorithm on 50bp regions flanking singleton insertion and deletion breakpoints, we were able to re-discover the sequence motifs of commonly disrupted binding sites, including the motifs of PRDM9 and NRF1 (Supplemental Table S9). In contrast, genome-wide, the motifs discovered flanking these variants were more likely to be simple repeats and other low complexity sequences that did not match known TFBS motifs, suggesting that processes other than transcription factor binding drive DNA breakage outside of active regulatory sites.

Genomic instability at spermatogonial TFBSs impacts enhancers active in neural development

Since many regulatory regions of the genome are active across a variety of cell types (Andersson et al. 2014), mutation at TFBSs in spermatogonia might disrupt gene regulation in other tissues. The developing brain is of particular interest, given reports of increased SV burdens in neurodevelopmental disorders (Girirajan et al. 2011; Leppa et al. 2016; Collins et al. 2017). We classified developmentally active human brain enhancers (distal regulatory elements) supported by neocortical ATAC-Seq data (de la Torre-Ubieta et al. 2018) according to whether they were either active (10,888 brain enhancers) or inactive in the male germline (26,162 brain enhancers). We then calculated the odds ratio of a singleton mutation affecting a brain enhancer which is also *active* in spermatogonia, relative to a brain enhancer which is *inactive* in spermatogonia. For DDD singleton deletion breakpoints, the odds ratio was 6.82 (95% CI = [5.34,8.71]), and for a singleton gnomAD insertion (5-20 bp), it was 4.69 (95% CI = [4.46,4.93]). This suggests that activity in spermatogonia greatly predisposes a brain enhancer to DNA damage, and this damage manifests in enhancers that share activity with the male germline (Figure 7). Brain enhancers that are shared with spermatogonia are, on average, more accessible in the developing brain than those that are inactive in the germline (the median “mean of normalized counts” for the two types of brain enhancers were 104.8 and 54.1, respectively; Wilcoxon test W = 197340000, p-value < 2.2e-16), suggesting a link between enhancer activity, the sharing of enhancers across tissues and propensity to mutation. The subset of brain enhancers which overlapped spermatogonial active sites were not enriched for specific motifs, and the number of motif sites for each motif family were highly correlated between brain and spermatogonia (Spearman's rho = 0.95, $p < 10^{-}$

[15]). That is, the propensity to mutation does not appear to be driven by an enrichment of specific motif families in brain enhancers.

Spermatogonia accessible TFBS motifs incur increased rates of disruption

We cannot exclude a small contribution of the TFBS sequence itself on the predisposition to mutation (Kondrashov and Rogozin 2004), but our data suggest that TF binding is a major driver of insertion and deletion mutation in the human germline. This is supported by the fact that we see an increase of disruption of brain enhancers if they are active in spermatogonia (Figure 7) and, more generally, an increase in the mutational load for sites that are active across other somatic tissue if binding also occurs in the germline (Supplemental Table S10). In addition, control motif sites (representing the same TFBS but located outside of ATAC-Seq peaks) are subject to lower rates of mutation compared to motifs within spermatogonial ATAC-Seq peaks (Figure 6c). Motifs within peaks carry, on average, 73% more mutations than their control counterparts, and for the most highly disrupted motifs, the discrepancy between active and control motifs is even larger. For example, PRDM9 motifs are 3.4-fold, HINFP 2.9-fold and NRF1 motifs 2.6-fold more disrupted if they are active in spermatogonia, relative to spermatogonia inactive motifs. We note that this increase in disruption is likely to be a conservative estimate since some control sites may be bound at time points in the germline that our ATAC-Seq data cannot ascertain.

Since the X chromosome spends only one third of its time in males - the sex with the higher number of germ cell divisions - a depletion of mutations on the X chromosome is expected for a male-biased mutational process. We find the X chromosome to be

strongly depleted for short singleton gnomAD insertions (5-20 bp), with a ratio of X to autosome variants per uniquely mappable site of 0.78 (Supplemental Table S11). However, we note that, despite the overall reduced rate of insertions on the X, ATAC-Seq peaks on the X are still subject to increased rates of insertions compared to genome-wide expectations, suggesting that the inferred effects of protein-binding on mutation are larger than the reduction in mutation due to X-linkage (38 observed insertions in X-linked ATAC-Seq peaks, whereas 11 were expected; $p < 10^{-4}$).

To test which candidate genomic feature most reliably predicts DNA damage, we used random forest regression to model the rate of singleton variants within 5 kb genomic windows, based on their overlap with spermatogonial TFBSs, ssDNA sites, LD-based hotspots, average GC content, mappability, gene density, replication time as well as various repeat families (LTRs, SINEs, LINEs and simple repeats). In models of genome-wide short insertion rates or deletion breakpoint rates, measures of replication timing and GC content were important predictors of mutation load as expected (Supplemental Fig. S9). Mappability was an important factor for predicting mutation rates genome-wide, perhaps reflecting the association between segmentally duplicated (low mappability) regions and rapid structural evolution, or perhaps suggesting that a fraction of variants may be erroneously called in the gnomAD dataset. (Only regions with high mappability were included in our more detailed analyses of TFBSs (Figures 3-7 and Supplemental Fig. S7)). However, spermatogonial ATAC-Seq derived TFBSs contributed additional predictive power to the models, even at the scale of the entire genome. The same TFBSs appear to be somewhat more important features in models that specifically predict damage at active brain enhancers (Supplemental Fig. S9). Genome-wide, deletion breakpoints

and 5-20 bp insertions were enriched in early replicating DNA (Spearman's rank correlation with replication timing: rho = 0.08, $p < 10^{-15}$ and rho = 0.07, $p < 10^{-15}$, respectively). In contrast, the presence of repeat elements had almost no impact in predicting either short insertion or deletion breakpoint rates (Supplemental Fig. S9). We conclude that germline active regulatory sites, through their occupancy by DNA binding factors, make a substantial contribution to genome-wide *de novo* structural variant rates, independent of other genomic features.

Discussion

We have demonstrated enrichments of rare and *de novo* SV breakpoints at spermatogonial regulatory sites defined by ATAC-Seq, suggesting that these sites suffer high rates of DSBs in the male germline. The same sites show unusual parallel enrichments for short variants, and particularly 5-20bp insertions. No TFBS family examined was found to be depleted for these rare variants, suggesting that many different TFBSs active in spermatogonia are prone to higher mutational loads. These loads appear to be positively correlated with the levels of chromatin accessibility/nucleosome disruption (ATAC-Seq peak binding strength) and the number of factors predicted to bind within the region. Sites bound by PRDM9, NRF1 and HINFP incur the highest levels of disruption, but 11% of 167 TF families examined showed evidence for significantly elevated mutation rates. These results have implications for the evolution of binding site patterns within regulatory regions, and for disrupted regulation in somatic tissues.

Homotypic clusters of TFBSs are a pervasive feature of both invertebrate and vertebrate genomes, and have long been known to be a common feature of human promoter and enhancer regions (Gotea et al. 2010). Various adaptive hypotheses have been proposed for the presence of such clusters such that they provide functional redundancy within a regulatory region, enable the diffusion of TF binding across a region, and allow cooperative DNA binding of TF molecules (Gotea et al. 2010). More recently it has been suggested that homotypic TFBS clusters may also contribute to phase separation and the compartmentalisation of the nucleus (Kribelbauer et al. 2019). Similarly, the clustered patterns of CTCF sites in the genome have been ascribed critical roles in chromatin architecture and regulation, particularly at regulatory domain boundaries. However, these boundary regions have been shown to exhibit genome instability (Kaiser and Semple 2018) and recurrently acquire new CTCF binding sites in dynamically evolving clusters (Kentepozidou et al. 2020). The data presented here suggest that binding site clusters may arise solely as a selectively neutral consequence of the unusual mutational loads at germline TFBSs, with clusters maintained by recurrent DNA damage and misrepair.

We observe significant enrichments of both large SV breakpoints and small insertions together at spermatogonial TFBSs. This parallel enrichment of both types of mutation may originate from DNA breakage, followed by misrepair, conceivably via a pathway such as non-allelic homologous recombination (NAHR). It is known that NAHR can create large insertions and deletions (Kim et al. 2016), and PRDM9 activity is implicated in certain developmental disorders arising via NAHR (McVean 2007; Myers et al. 2008; Berg et al. 2010). For example, the locations of PRDM9 binding hotspots coincide with recurrent SV breakpoints causing Charcot-Marie-Tooth

417 disease, and Hunter and Potocki-Lupski/Smith-Magenis syndromes (Pratto et al.
418 2014). It is possible that the sequence similarity at TFBSs scattered across the genome
419 may make them particularly prone to NAHR. However, we note that the sequence
420 similarity between the low copy repeat units, known to be involved in NAHR, is
421 usually of the size of several kb (Gu et al. 2008), rather than sequences on the scale of
422 TFBSs. The NHEJ pathway can also lead to short insertions after DNA breakage,
423 usually in G0 and G1 phases of the cell cycle. Indeed, NHEJ is the most common
424 repair pathway of DSBs in mammals and it is typically error prone (van Gent et al.
425 2001; Lieber et al. 2003). During NHEJ, double-strand break ends are resected to
426 form single-stranded overhangs, but when pairing occurs between the tips of the
427 overhangs, sequences near the breakpoints will often be duplicated (Rodgers and
428 McVey 2016). Interestingly and consistent with our results based on ultra-rare
429 sequence variants, two previous studies using human–chimpanzee–macaque multiple
430 alignments have shown that high numbers of short insertions have occurred in the
431 human lineage (Kvikstad et al. 2007; Messer and Arndt 2007), and both conclude that
432 these insertions preferentially take place in the male germline, evidenced by
433 decreased mutation rates on the X chromosome, with similar observations in rodents
434 (Makova et al. 2004).
435
436 The data presented here suggest that different DNA binding proteins differ widely in
437 their impact on mutation rates. The two proteins with the largest impacts, NRF1 and
438 PRDM9, are both highly expressed in testis, revealing a possible link between the
439 expression level of a gene encoding a DNA binding protein and the propensity for
440 breakage or inefficient repair at the sites the protein binds. Incidentally, NRF1 has a
441 pLI score of 0.999, indicating that it is extremely loss-of function intolerant and

crucial for the organism's functioning (Karczewski et al. 2020). A previous study (Montgomery et al. 2013), using 1K genomes polymorphism data, failed to find an increase in indels at PRDM9 motifs genome-wide. This highlights the importance of using ATAC-Seq data to confine the search for motifs to germline active sites only, combined with singleton variants from large-scale sequencing studies as a more powerful strategy to explore fine scale mutational patterns.

Although studies of coding sequences, such as the DDD (Deciphering Developmental Disorders Study 2015), have revealed many of the genes disrupted in developmental disorders, more than half of cases lack a putatively causal variant (Mcrae et al. 2017), stimulating interest in the noncoding remainder of the genome, and particularly regulatory regions active in development. Limited sequencing data, covering a fraction of human regulatory regions, suggests that *de novo* mutations are enriched in these regions and are therefore likely to contribute to neurodevelopmental disorders at some level (Short et al. 2018; Gerrard et al. 2020). However, there appear to be very few, if any, individual regulatory elements recurrently mutated across multiple cases to cause neurodevelopmental disorders with a dominant mechanism (Short et al. 2018). The data presented here suggests a potential solution to this paradox, where combinations of mutations at multiple regulatory regions may underlie a disease phenotype. The frequency of such combinations is expected to be many times higher if they involve regulatory regions bound by factors such as NRF1. In such cases, an entire class of sites, rather than an individual site, is subject to recurrent mutation.

Methods

Identification of spermatogonial binding sites

Samples of testicular tissue were obtained from three patients undergoing orchiectomy with total processing completed within ~5-7 hours of explant. Tissue was obtained after informed consent through the Lothian NRS BioResource, and the study was approved by NHS Lothian (Lothian R&D Project Number 2015/0370TB). Tissue samples were disaggregated into cells, and cells were labelled with phycoerythrin (PE)-conjugated antibody against the cell surface marker FGFR3 (FAB766P, clone 136334, R&D systems). Spermatogonial cells were isolated using a FACSAria II cell sorter (BD bio- sciences) based on PE fluorescence and cell shape, according to Forward/Side Scatter. Isolated cells were subjected to ATAC-seq using the protocol and reagents described in (Buenrostro et al. 2013), followed by paired-end sequencing on Illumina HiSeq4000 (75 bp read length). We combined reads from separate sequencing runs into three biological replicates, based on origin and morphological appearance of the FACS sorted cells. Replicate 1: combined sequencing runs H.5.1 and H.5.4; a non-cancer patient; large cells, high side scatter; 58,000 and 42,000 cells, respectively. Replicate 2: combined sequencing runs H.5.2 and H.5.5; the same non-cancer patient as Replicate 1; large cells; 36,000 and 23,000 cells, respectively. Replicate 3: combined sequencing runs H.7.3 and H.10.2; normal tissue from cancer patients; large cells; 69,000 and 24,000 cells, respectively. ATAC-seq raw reads were trimmed to remove any retained adaptor sequences using cutadapt (Martin 2011) with parameters *-n3 --format=fastq --overlap=3 -g GAGATGTGTATAAGAGACAG -g CAGATGTGTATAAGAGACAG –a CTGTCTCTTATACACATCTG -a CTGTCTCTTATACACATCTC*. Reads were aligned to the GRCh38/hg38 genome assembly with Bowtie2 (Langmead and Salzberg 2012) in paired-end mode, limiting the insert size to 4 kb. Any reads with quality score < 30

were discarded. Paired end reads were converted to fragments using bedtools bamtobed -bedpe, followed by extraction of the most 5' and 3' coordinates of each pair. PCR duplicates were removed by retaining only one instance of a fragment with identical coordinates within a sample. Fragments overlapping with the regions previously blacklisted as mitochondrial homologs (Buenrostro et al. 2013) were discarded. Peaks were identified from short fragments of <= 100 bp (Supplemental Fig. S1), thought to arise due to transposition events around transcription factor binding sites – and distinct from fragments spanning the larger nucleosomes (characterized by a ~200 bp periodicity) (Buenrostro et al. 2013). Peaks were called from short fragments using macs2 callpeak (Zhang et al. 2008) with the following parameters: -B -q 0.01 -f BAMPE --nomodel --nolambda --keep-dup auto --call-summits. The clustering of ATAC-Seq peaks near transcription start sites and promoters was assessed using the ChIPseeker R package (Yu et al. 2015).

For the downstream mutation analyses, ATAC-Seq peaks from Replicates 1 and 2 (the non-cancer patient) were merged, creating a single peak set. This dataset also formed the basis for the footprinting analysis, which used, as input, the combined short sequencing fragments of Replicates 1 and 2, running "rgt-hint footprinting" with --atac-seq and --bias-correction, followed by "rgt-motifanalysis matching" with the option --remove-strand-duplicates (Li et al. 2019). Input motifs were the 579 position weight matrixes (PWMs) of the japspar vertebrate database (Sandelin et al. 2004) as well as the 13-mer PRDM9 motif "CCNCCNTNNCCNC" (Myers et al. 2010) which was also provided as a PWM. The tissue donor for Replicates 1 and 2 was a carrier of the most common (European) alleles of PRDM9, which was confirmed by investigating his allelic state at the SNP (rs6889665) identified by Hinch et al. (2011);

this SNP was covered by our ATAC-Seq by 10 reads, all of which were "T". Accordingly, we assume that the donor is a carrier of the A and/or B allele of PRDM9 (both of which bind the same DNA motif), and the search for the 13-mer PRDM9 motif in this patient's ATAC-Seq data can be used as a proxy for PRDM9 binding in European populations. In addition, Replicate 3 was processed in the same way as the combined Replicates 1 and 2 and served as a positive control to assess the genome-wide enrichment of mutations at spermatogonial accessible sites (Supplemental Fig. S3).

Jaspar input motifs are often highly similar, resulting in multiple binding proteins being identified by the rgt-hint pipeline to bind at the same ATAC-Seq footprint; this is biologically implausible (since only one protein is likely to occupy a given site), and we clustered motifs by similarity, using the default parameters of the PWMclus CCAT package (Jiang and Singh 2014). This resulted in a set of 167 motif families of similar binding motifs (Supplemental Table S6). Using bedtools (Quinlan and Hall 2010), we merged overlapping binding sites that belonged to motifs of the same family (thus calling them only once), and we also merged palindromic binding sites called on both strands. Since PRDM9 is known to leave a characteristic histone methylation mark on bound DNA (Grey et al. 2011; Powers et al. 2016), we intersected the PRDM9 motif sites with testis-derived H3K4me3 marks (called in an PRDM9 A/B heterozygous individuals) from Pratto et al. (2014). This resulted in a stringent set of PRDM9 sites, which were both located in ATAC-Seq footprints and also carried the H3K4me3 mark in human testis. ATAC-Seq-defined PRDM9 sites showed moderate overlap with DMC1-bound ssDNA sites (Pratto et al. 2014) as well as recombination HSs (Myers et al. 2005), which may reflect the fact that most cells

in our experiments are likely to be pre-meiotic: only 10 and 11% of PRDM9 sites were within 500 bp of a ssDNA peak and a recombination HS, respectively, whereas 44% of DMC1-bound sites overlap with LD-defined HSs. However, we find that stronger ssDNA peaks are more likely to be near a PRDM9-binding site (Supplemental Fig. S10).

Comparisons between ATAC-Seq datasets

Using the same procedure as described above, we processed raw ATAC-Seq reads from previously published datasets in order to call MACS2 peaks from short sequencing fragments. Datasets included ATAC-Seq reads from the germinal zone and cortical plate of the developing brain (SRR6208926, SRR6208927, SRR6208938, SRR6208943) (de la Torre-Ubieta et al. 2018), ATAC-Seq experiments of KIT+ spermatogonia (sra accessions SRR7905001 and SRR7905002) (Guo et al. 2018), SSEA4+ spermatogonia (SRR5099531, SRR5099532, SRR5099533, SRR5099534) (Guo et al. 2017) and ESC cells (SRR5099535 and SRR5099536) (Guo et al. 2017). Adapter sequences within raw sequencing data were identified using bbmerge.sh of bbmap (https://sourceforge.net/projects/bbmap/) and removed using cutadapt (Martin 2011), as above. Encode ATAC-Seq datasets (Encode Project Consortium 2012; Davis et al. 2018) (Liver: ENCFF628MCV, Ovary: ENCFF780JBA, Spleen: ENCFF294ZCT, Testis: ENCFF048IOT, Transverse Colon: ENCFF377DAO) were downloaded as bam files, converted to BEDPE format, and short fragments were identified for peak calling.

BedGraph files (from the MACS2 output), describing the fragment pileup, were converted to bigwig format using bedGraphToBigWig and uploaded to the Galaxy server at https://usegalaxy.eu/(Afgan et al. 2018). DeepTools2's multiBigwigSummary (with default parameters) and plotCorrelation (with parameters –skipZeros –removeOutliers) (Ramirez et al. 2016) were used to create a heatmap of ATAC-Seq signals in the different tissues and datasets. The KIT+ and SSEA4+ spermatogonial ATAC-Seq datasets of Guo et al. (2017) and Guo et al. (2018) were further used to perform footprinting and motif matching analyses (Li et al. 2019) as described above for the FGFR3-positive cells. Peaks and motif sites that are accessible in the developing brain but not in spermatogonia were identified using bedtools intersect (Quinlan and Hall 2010).

Structural Variant Breakpoint data

Large SVs, identified by high-density arrayCGH, or a combination of arrayCGH + exome sequencing, were extracted from a cohort of 9,625 DDD patients, using variant calling procedures as described in (Deciphering Developmental Disorders Study 2015). We filtered the DDD variants to only keep variants which fulfilled the following criteria: a CNsolidate wscore >= 0.468, a callp < 0.01 and a mean log2 ratio of < -0.41 for deletions and 0.36 for duplications; CIFER "false positives" were removed. Singleton variants were identified as being annotated as "novel" by the DDD release, only seen once among the DDD patients, and not seen in the dgv (MacDonald et al. 2014) and gnomAD V.2 (Collins et al. 2020) structural variant datasets (80% reciprocal overlap criterion). Since there are 9,625 patients in the DDD dataset, the gnomAD V.2 dataset contains SVs from 10,738 genomes and the dgv

591 contains SVs from 29,084 individuals, this puts an upper limit of the frequency of
592 carriers of a singleton variant at ~ 0.002%. Breakpoints were identified as the 5' and
593 3' coordinates of SVs, resulting in 13,406 singleton deletion and 3,406 duplication
594 breakpoints; the resolution of the breakpoints was such that the median and mean
595 confidence intervals were 300 bp and 12 kb, respectively. Further, we identified
596 11,962 "common" deletion variants in the DDD dataset, which had a minimum
597 variant frequency of 1% in the consensus CNV dataset as described by the DDD
598 study (2015), i.e. pooled CNV datasets of Conrad et al. (2010), the Genomes Project
599 Consortium (2010), the Wellcome Trust Case Control (2010) and the DDD normal
600 controls. We used the 80% reciprocal overlap criterion and grouped common variants
601 using the bedmap options --echo-map --fraction-both 0.8, followed by bedops --merge
602 (Neph et al. 2012). The breakpoints of common variants are thus the outermost
603 coordinates of all SVs that are collapsed into a given variant. The overlap of such
604 "common" breakpoints with ATAC-Seq peaks was assessed independently of SV
605 allele frequencies, i.e. a group of common SVs contributed two breakpoints to the
606 analysis.
607 We also identified a set of singleton CNVs called with the Manta algorithm (Chen et
608 al. 2016) from the gnomAD V.2 database (Collins et al. 2020) (80% reciprocal
609 overlap criterion with gnomAD V.2, dgv and DDD variants), resulting in a set of
610 73,063 deletion and 15,419 duplication breakpoints seen in ~ 0.002% of individuals
611 but called with a different approach compared to the DDD. Common deletions and
612 duplications ($p >= 0.05$) were also extracted from the gnomAD V.2 dataset; these
613 variants had also been called with the Manta algorithm and included 5,954 deletion
614 and 1,586 duplication breakpoint sites.
615

Indels and SNP data

The recently released gnomAD V.3 variants (indels and SNPs) were downloaded from https://gnomad.broadinstitute.org/. Only variants that passed all filters were kept (filtering using VCFtools --remove-filtered-all). Multiallelic variants were split using bcftools norm, and bcftools norm --IndelGap 2 was applied to indels, to allow only variants to pass that were separated by at least 2 bp. Singleton variants were defined as having an allele count of one, and the allele number was >=100,000, i.e. the allele frequency of singletons was $p <= 0.001\%$.

We subdivided gnomAD indels into singleton insertions and deletions of different sizes: 1-2 bp (most commonly arising due to replication slippage) and those 5-20 bp (arising due to other mechanisms of DNA instability and within the size range reliably detected by short-read sequencing). To speed up simulations and allow for easy comparison between categories of variants, all classes of InDels and single nucleotide variants were down-sampled to 650,000 variants each.

A total of 854,409 *de novo* SNPs and indels were compiled from three different sources, lifted over to the hg38 assembly using the UCSC liftOver tool as required. First, we downloaded variants from http://denovo-db.gs.washington.edu/, including only samples from whole genome sequencing studies (Michaelson et al. 2012; Ramu et al. 2013; Genome of the Netherlands 2014; Besenbacher et al. 2015; Turner et al. 2016; Yuen et al. 2016; Jonsson et al. 2017; RK et al. 2017; Turner et al. 2017; Werling et al. 2018), which included a total of 404,238 variants from 4,560 samples. Additional samples, which were not already included in the denovo-db dataset, were downloaded from the MSSNG database (https://research.mss.ng/), version

2019/10/16, which added 2,243 samples and 215,044 *de novo* mutations. A third source of *de novo* variants came from(An et al. 2018) - 3,805 samples and 255,107 mutations.

Circular Permutation

To obtain a genome-wide estimate of enrichment of overlap between genomic features (e.g. TFBSs and mutations), we performed circular permutations using the Bioconductor regioneR package in R (Gel et al. 2016). We used the permTest() function with parameters ntimes=10000, randomize.function=circularRandomizeRegions, evaluate.function=numOverlaps, genome=hg38_masked, alternative="auto", where hg38_masked = getBSgenome("BSgenome.Hsapiens.UCSC.hg38.masked"). This test evaluates the number of overlaps observed between two sets of genomic features, given their order on the chromosome and the distance between features, i.e. taking their degree of clustering into account. At each iteration, all positions are shifted by the same randomly generated distance on conceptually "circularized" chromosomes, in effect “spinning” the position of features on each chromosome, while excluding masked regions (i.e. unmappable, repetitive and low-complexity segments). The statistical output is a z-score, which is defined as the distance between the expected number of overlaps (the distribution over 10,000 permutations) and the observed one, measured in terms of standard deviations. Shifted Z-score analysis can further indicate whether a given overlap pattern is caused by the precise locations of two feature sets or by broader, regional effects. One set of features is shifted in either direction from their original positions by a number of bases, so that the numbers of overlaps and the

degree to which the z-score changes in response to the shift can be tested. A sharp peak of shifted z-scores around the zero-coordinate indicates a precise overlap between features, whereas a flat profile may indicate overlaps attributable to regional effects, as is seen for features that tend to co-occur in regions with similar base composition or gene density.

For permutations involving SVs, we used the two breakpoints of each SV, and assessed the overlap of breakpoints with another feature of interest (i.e. ATAC-Seq sites), treating each breakpoint separately.

Circular permutations in regioneR (Gel et al. 2016) were also used assess the mean distance between ATAC-Seq peaks and deletion breakpoints, for common and singleton variants separately.

Simple permutations

Spermatogonial binding sites were randomly permuted, using the “bedtools shuffle” command (Quinlan and Hall 2010), with the parameter “-noOverlapping” and excluding assembly gaps with “-excl hg38.gap.bed” (downloaded from https://genome.ucsc.edu/cgi-bin/hgTables). In each of 10,000 simulations, we assessed the overlap of the permuted binding sites with short insertions and deletion breakpoints, respectively. This resulted in an expected distribution of overlap, given a random positioning of binding sites across the genome. This distribution was compared to the observed number of overlaps (using bedtools intersect), and the p-value was defined as the percentage of simulated overlaps that was larger than the observed overlap. This permutation framework does not take into account the spacing

and clustered nature of binding sites, and does not allow for an assessment of the precision of overlap between features. In this study, we favor the more conservative measures of significance provided by the circular permutation strategy.

Brain enhancer data

Active brain enhancers came from de la Torre-Ubieta et al. (2018). Specifically, we used the 37,050 brain enhancers which showed differential accessibility in the germinal zone versus the cortical plate, reflecting activity in the developing brain (de la Torre-Ubieta et al. 2018). Next, we identified brain enhancers that were also active during the male germline formation, i.e. overlapping the spermatogonial ATAC-Seq peaks. To correct for the variable size of the brain active enhancers, we took the midpoints of each enhancer plus/minus 500 bp on either side, and intersected these sites with the ATAC-Seq peaks using bedtools intersect (Quinlan and Hall 2010), thus classifying brain enhancers as spermatogonial “active” or “inactive”. Next, we intersected these two categories of brain enhancers with the DDD breakpoint and gnomAD insertion dataset, respectively, to further classify them as “disrupted” by a singleton variant or “intact”. An odds ratio was calculated as

$$OR = (A/(B - A))/(C/(D - C))$$

With confidence intervals

$$CI\_lower = \exp(\log(OR) - 1.96 * \sqrt{1/A + 1/(B-A) + 1/C + 1/(D-C)})$$

$$CI\_higher = \exp(\log(OR) + 1.96 * \sqrt{1/A + 1/(B-A) + 1/C + 1/(D-C)})$$

where:

A = Disrupted, sperm active

B = All sperm active

C = Disrupted, sperm inactive

D = All sperm inactive

To analyse the enrichment of short InDels and SNPs around TFBSs and brain enhancers, we only considered genomic regions with unique mappability in >= 95% of the region, using the bedmap option --bases-uniq-f (Neph et al. 2012) and the mappability file hg38_umap24 (Karimzadeh et al. 2018), converted to bedmap format.

Random Forest Regression

To compare the effects of chromatin state on mutation rates, we performed random forest regression with 200 trees, modelling the outcome variables “singleton breakpoints” and “singleton insertions (5-20 bp)”, from the DDD and gnomAD V.3 respectively, within 5-kb wide genomic windows. Predictor variables included “spermatogonial TFBS count”, “ssDNA overlap” (from Pratto et al.(2014)), “recombination HS overlap” (from The 1000 Genomes Project et al. (2015)), “GC-content”, “Replication timing” (average of Wavelet-smooth signal in 1-kb bins of 15 encode tissues, downloaded from http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/), “Gene density”, “Mappability” (proportion of sites in each window with an umap24 score of 1), and the overlap with “LTRs”, “SINEs”, “LINEs” and “Simple Repeats” (downloaded from the Table Browser at https://genome.ucsc.edu/).

In a smaller model, we subsetted the dataset to only include 5-kb bins that also overlap active brain enhancers (de la Torre-Ubieta et al. 2018), then ran the random forest regression model to predict mutation rates within genomic regions that contain active brain enhancers.

Motif discovery in singleton insertion sites

In order to find sequence motifs within the 5-20 bp singleton insertion sites from gnomAD V.3, without prior assumptions, we extracted the fasta sequence for insertions that fell within 10 bp of the top 10 disrupted motif families (motif families 992, 193, 796, 907, 579, 825, 984, 171, 991). We ran the meme.4.11 motif discovery algorithm (Bailey et al. 2009) with "-nmotifs 1" on the inserted sequences. This allowed us to compare the sequence motif of the disrupted TFBSs to any recurrent motif found within the inserted sequences.

Control Motif sites

Using default search criteria, the FIMO algorithm (Grant et al. 2011) was run on the repeat masked hg38 genome sequence (hg38.fa.masked, downloaded from https://genome.ucsc.edu/ in March 2020), searching the whole genome for the 579 input Jaspar motifs and the 13-mer PRDM9 motif. As with active binding sites, motif matches belonging to the same motif family were merged and reported as a single motif match per family, and only regions with unique umap24 mappabilities for >= 95% of sites were kept; motifs that overlapped with spermatogonial ATAC-Seq peaks were excluded. Next, these "control" motif sites were down-sampled to 10,000 per

766 motif family (using bedtools sample (Quinlan and Hall 2010)); circular permutations
767 were performed to compare the observed to expected overlap of the control motif sites
768 (plus/minus 10 bp) with the gnomAD singleton insertions of 5-20 bp.
769
770 The FIMO predicted control sites were also used to assess the degree of "clustering"
771 of motifs at spermatogonia active sites. For this purpose, we intersected the FIMO
772 motifs with a) spermatogonial ATAC-Seq sites and b) ENCODE Master regulatory
773 sites downloaded from https://genome.ucsc.edu/ (DNaseI hypersensitivity derived
774 from assays in 95 cell types). For each of the 167 motif families, we calculated the
775 median distance (in basepairs) from a motif located within the active regulatory
776 region to the nearest FIMO motif of the same type. Accordingly, the ratio of the
777 median distance between motif sites (ENCODE/spermatogonia) was larger than one if
778 motifs at spermatogonial sites were, on average, closer to each other than motifs near
779 ENCODE sites, and we used this ratio as a measure of motif clustering. When
780 correlating the IFE with the degree of motif clustering (Figure 6), we thus largely
781 correct for base compositional biases near active sites (which impact mutation rates –
782 Supplemental Fig. S9) as well as the effects of historical selection on the clustering of
783 motifs near genes, i.e. shorter inter-motif distances in spermatogonia indicate that
784 these sites have specifically high levels of motif density in spermatogonia, beyond the
785 levels expected for binding sites in general.
786
787 **Data access**
788
789 Raw sequencing data generated in this study have been submitted to the European
790 Genome-phenome Archive (EGA) (accession number EGAS00001005366), and

ATAC-Seq peak files are available at Edinburgh DataShare (https://doi.org/10.7488/ds/3053).

Competing interest statement

The authors have no competing interests to declare.

Acknowledgements

We thank all donors for their participation in genetic research. In particular, we thank the DDD families, study clinicians, research nurses and clinical scientists in the recruiting centres; the Genome Aggregation Database (http://gnomad.broadinstitute.org/), MSSNG (https://www.mss.ng/) and denovo-db (http://denovo-db.gs.washington.edu/denovo-db/) for making their data available. We are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, E. Wijsman). We appreciate obtaining access to genetic data on SFARI Base. Approved researchers can obtain the SSC population dataset described in this study by applying at https://base.sfari.org.

This work was supported by MRC Human Genetics Unit core funding programme grants MC_UU_00007/11, MC_UU_00007/2 and MC_UU_00007/16.

We thank Elisabeth Freyer for assistance with the FAC sorting and Wendy A. Bickmore for useful comments to the manuscript.

Author Contributions

V.B.K. and C.A.S. conceived the project, interpreted the results and wrote the manuscript. M.S.T. designed the experiments and managed the acquisition of samples. L.T., Y.K., F.S. and M.M. performed the experiments, L.T. processed raw data. V.B.K. performed the analyses. D.D.D. and D.R.F. provided data. D.R.F. helped with the interpretation of the results and provided critical scientific inputs.

Figure Legends

Figure 1: Locally elevated structural variation rates at spermatogonial regulatory sites. SV breakpoint count (a, b) and circular permutation shifted Z-scores (c, d) of deletion breakpoints in the DDD cohort, centred around the midpoints of spermatogonial ATAC-Seq peaks. “Singletons” are breakpoints of deletions with a frequency of ~ 0.002% across population samples; “common” variants are seen at a frequency of at least 1% in the DDD consensus dataset (see main text); permutation p-values indicate significant enrichment for both types of variants at ATAC-Seq peaks ($p < 10^{-5}$ in each case) (e, f) Circular permutation shifted Z-scores of gnomAD deletion breakpoints, centred around spermatogonial ATAC-Seq peaks. “Singletons” are breakpoints of deletions with a frequency of ~ 0.002% across population samples; “common” variants are seen at a frequency of at least 5% in the gnomAD V.2 dataset.

Permutation p-values indicate significant enrichment for singleton breakpoints ($p < 10^{-5}$), and a significant depletion for common variants ($p < 0.01$).

Figure 2: Increased rates of short insertions focussed on spermatogonial binding sites. Insertion count (**a, b**) and Shifted Z-scores (**d, e**) of gnomAD singleton and common insertions (5-20 bp), centred around spermatogonial TFBSs. Singletons are seen only once in the gnomAD V.3 dataset (allele frequency <= 0.001%) and are significantly enriched at binding sites (p < 10-4); common variants have an allele frequency of at least 5% within gnomAD V.3 and are significantly depleted at binding sites (p < 10-4).

Figure 3: Parallel enrichments of short variants and SV breakpoints at spermatogonial binding sites. Circular permutation results are based on 10,000 permutations; results for singleton variants and de novo mutation are shown. The Y axis shows the ratio of observed over expected variant counts. Mutation categories with significant enrichment are indicated by asterisks (*** indicating p < 0.001). The type of variant tested and the total number of observed variants overlapping spermatogonial TFBSs are indicated below each bar.

Figure 4: Binding factors associated with the highest rates of mutation at spermatogonial binding sites. Plots are centred on the binding sites of a given motif family inside ATAC-Seq footprints. (**a**) Singleton and (**b**) common deletion breakpoints in the DDD cohort; singletons are breakpoints of deletions with a frequency of ~ 0.002 across population samples; common variants are seen at a frequency of at least 1% in the DDD consensus dataset. (**c**) Singleton and (**d)** common insertions (5-20 bp) in the gnomAD dataset. Singletons are seen only once in gnomAD V.3 (allele frequency <= 0.001%), and common variants have an allele frequency of at least 5% within gnomAD V.3. Only 10 kb regions around TFBSs with

>=95% unique mappability (umap24 scores) were included. The top 5 disrupted motifs are shown, listed in order of enrichment of singleton variants in the circular permutations (all enrichments of singletons are associated with p-values < 10^{-4}).

Figure 5: Elevated singleton insertion rates at PRDM9 and NRF1 binding sites contrast with other short variant classes. All gnomAD variants have been down-sampled to a total of 650,000 variants per analysis, making the Y axes directly comparable; individual bins are 5 bp in size. Only regions around TFBSs with >=95% unique mappability (umap24 scores) were included.

Figure 6: Insertions at spermatogonial TFBSs generate motif clusters in the genome. a) Jaspar database sequence motifs identified in the footprints of spermatogonial ATAC-Seq peaks (left) and the motifs identified in the singleton insertions (5-20 bp) (right). The number of insertion sites (N) that were chosen by MEME to construct the motif are shown on the right. **b)** For each motif family, we plot the insertion fold enrichment (IFE) on the X axis and the degree of spermatogonial motif clustering on the Y axis; the least square regression line is indicated in blue. Motif clustering is measured as the distance to the nearest motif at spermatogonial active sites, relative to the distance for motifs at ENCODE active sites. c) The insertion fold enrichment (IFE) is contrasted between FIMO control motif sites (left) and spermatogonial active motif sites (right); the Wilcoxon Test was performed to compare the IFE at the two classes of sites.

Figure 7: Neural enhancers with activity in spermatogonia suffer elevated mutation rates. **a)** Singleton DDD deletion breakpoint and **b)** singleton gnomAD insertion (5-20 bp) count around brain active enhancers. Enhancers were classified as being also active in spermatogonia (red) or inactive in spermatogonia (blue). Plotted is the average number of variants per brain enhancer - in 5 kb windows or 100 bp windows,

respectively. In **b**, only 10 kb regions around enhancers with >=95% unique mappability (umap24 scores) were included (3,409 brain enhancers that are inactive in spermatogonia and 1,029 that are active).

References

Afek A, Shi H, Rangadurai A, Sahay H, Senitzki A, Xhani S, Fang M, Salinas R, Mielko Z, Pufall MA et al. 2020. DNA mismatches reveal conformational penalties in protein-DNA recognition. *Nature* **587**: 291-296.

Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Gruning BA et al. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* **46**: W537-W544.

Altemose N, Noor N, Bitoun E, Tumian A, Imbeault M, Chapman JR, Aricescu AR, Myers SR. 2017. A map of human PRDM9 binding provides evidence for novel behaviors of PRDM9 and other zinc-finger proteins in meiosis. *Elife* **6**.

An JY, Lin K, Zhu L, Werling DM, Dong S, Brand H, Wang HZ, Zhao X, Schwartz GB, Collins RL et al. 2018. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**.

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**: 455-461.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**: W202-208.

Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, de Massy B. 2010. PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. *Science* **327**: 836-840.

Berg IL, Neumann R, Lam KW, Sarbajna S, Odenthal-Hesse L, May CA, Jeffreys AJ. 2010. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nat Genet* **42**: 859-863.

Besenbacher S, Liu S, Izarzugaza JM, Grove J, Belling K, Bork-Jensen J, Huang S, Als TD, Li S, Yadav R et al. 2015. Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat Commun* **6**: 5969.

Besenbacher S, Sulem P, Helgason A, Helgason H, Kristjansson H, Jonasdottir A, Jonasdottir A, Magnusson OT, Thorsteinsdottir U, Masson G et al. 2016. Multi-nucleotide de novo Mutations in Humans. *PLoS Genet* **12**: e1006315.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**: 1213-1218.

Carlson J, Locke AE, Flickinger M, Zawistowski M, Levy S, Myers RM, Boehnke M, Kang HM, Scott LJ, Li JZ et al. 2018. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nat Commun* **9**.

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Kallberg M, Cox AJ, Kruglyak S, Saunders CT. 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**: 1220-1222.

Collins RL Brand H Karczewski KJ Zhao X Alföldi J Francioli LC Khera AV Lowther C Gauthier LD Wang H et al. 2020. A structural variation reference for medical and population genetics. *Nature* **581**: 444-451.

Collins RL, Brand H, Redin CE, Hanscom C, Antolik C, Stone MR, Glessner JT, Mason T, Pregno G, Dorrani N et al. 2017. Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol* **18**: 36.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P et al. 2010. Origins and functional impact of copy number variation in the human genome. *Nature* **464**: 704-712.

Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK et al. 2018. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* **46**: D794-D801.

de la Torre-Ubieta L, Stein JL, Won H, Opland CK, Liang D, Lu D, Geschwind DH. 2018. The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. *Cell* **172**: 289-304 e218.

Deciphering Developmental Disorders Study. 2015. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**: 223-228.

Encode Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

Gel B, Diez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R. 2016. regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* **32**: 289-291.

Genome of the Netherlands C. 2014. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46**: 818-825.

Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061-1073.

Gerrard DT, Berry AA, Jennings RE, Birket MJ, Zarrineh P, Garstang MG, Withey SL, Short P, Jimenez-Gancedo S, Firbas PN et al. 2020. Dynamic changes in the epigenomic landscape regulate human organogenesis and link to developmental disorders. *Nat Commun* **11**: 3920.

Girirajan S, Brkanac Z, Coe BP, Baker C, Vives L, Vu TH, Shafer N, Bernier R, Ferrero GB, Silengo M et al. 2011. Relative burden of large CNVs on a range of neurodevelopmental phenotypes. *PLoS Genet* **7**: e1002334.

Gonzalez-Perez A, Sabarinathan R, Lopez-Bigas N. 2019. Local Determinants of the Mutational Landscape of the Human Genome. *Cell* **177**: 101-114.

Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I. 2010. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res* **20**: 565-577.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**: 1017-1018.

Grey C, Barthes P, Chauveau-Le Friec G, Langa F, Baudat F, de Massy B. 2011. Mouse PRDM9 DNA-Binding Specificity Determines Sites of Histone H3 Lysine 4 Trimethylation for Initiation of Meiotic Recombination. *Plos Biol* **9**.

Gu W, Zhang F, Lupski JR. 2008. Mechanisms for human genomic rearrangements. *Pathogenetics* **1**: 4.

Guo J, Grow EJ, Mlcochova H, Maher GJ, Lindskog C, Nie X, Guo Y, Takei Y, Yun J, Cai L et al. 2018. The adult human testis transcriptional cell atlas. *Cell Res* **28**: 1141-1157.

Guo J, Grow EJ, Yi C, Mlcochova H, Maher GJ, Lindskog C, Murphy PJ, Wike CL, Carrell DT, Goriely A et al. 2017. Chromatin and Single-Cell RNA-Seq Profiling Reveal Dynamic Signaling and Metabolic Transitions during Human Spermatogonial Stem Cell Development. *Cell Stem Cell* **21**: 533-546 e536.

Halldorsson BV, Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, Gunnarsson B, Oddsson A, Halldorsson GH, Zink F et al. 2019. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**.

Hinch AG, Tandon A, Patterson N, Song YL, Rohland N, Palmer CD, Chen GK, Wang K, Buxbaum SG, Akylbekova EL et al. 2011. The landscape of recombination in African Americans. *Nature* **476**: 170-U167.

Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* **12**: 756-766.

Jiang P, Singh M. 2014. CCAT: Combinatorial Code Analysis Tool for transcriptional regulation. *Nucleic Acids Research* **42**: 2833-2847.

Jonsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT, Hjorleifsson KE, Eggertsson HP, Gudjonsson SA et al. 2017. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**: 519-522.

Kaiser VB, Semple CA. 2018. Chromatin loop anchors are associated with genome instability in cancer and recombination hotspots in the germline. *Genome Biol* **19**: 101.

Kaiser VB, Taylor MS, Semple CA. 2016. Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. *PLoS Genet* **12**: e1006207.

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**: 434-443.

Karimzadeh M, Ernst C, Kundaje A, Hoffman MM. 2018. Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Res* **46**: e120.

Kentepozidou E, Aitken SJ, Feig C, Stefflova K, Ibarra-Soria X, Odom DT, Roller M, Flicek P. 2020. Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. *Genome Biol* **21**: 5.

Kim S, Peterson SE, Jasin M, Keeney S. 2016. Mechanisms of germ line genome instability. *Semin Cell Dev Biol* **54**: 177-187.

Kondrashov AS, Rogozin IB. 2004. Context of deletions and insertions in human coding sequences. *Hum Mutat* **23**: 177-185.

Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, Sunyaev SR, McCarroll SA. 2012. Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am J Hum Genet* **91**: 1033-1040.

1029 Kribelbauer JF, Rastogi C, Bussemaker HJ, Mann RS. 2019. Low-Affinity Binding
1030 Sites and the Transcription Factor Specificity Paradox in Eukaryotes. *Annu*
1031 *Rev Cell Dev Biol* **35**: 357-379.
1032 Kvikstad EM, Tyekucheva S, Chiaromonte F, Makova KD. 2007. A macaque's-eye
1033 view of human insertions and deletions: differences in mechanisms. *PLoS*
1034 *Comput Biol* **3**: 1772-1782.
1035 Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat*
1036 *Methods* **9**: 357-359.
1037 Leppa VM, Kravitz SN, Martin CL, Andrieux J, Le Caignec C, Martin-Coignard D,
1038 DyBuncio C, Sanders SJ, Lowe JK, Cantor RM et al. 2016. Rare Inherited and
1039 De Novo CNVs Reveal Complex Contributions to ASD Risk in Multiplex
1040 Families. *Am J Hum Genet* **99**: 540-554.
1041 Levinson G, Gutman GA. 1987. High frequencies of short frameshifts in poly-CA/TG
1042 tandem repeats borne by bacteriophage M13 in Escherichia coli K-12. *Nucleic*
1043 *Acids Res* **15**: 5323-5338.
1044 Li C, Luscombe NM. 2020. Nucleosome positioning stability is a modulator of
1045 germline mutation rate variation across the human genome. *Nat Commun* **11**:
1046 1363.
1047 Li ZJ, Schulz MH, Look T, Begemann M, Zenke M, Costa IG. 2019. Identification of
1048 transcription factor binding sites using ATAC-seq. *Genome Biology* **20**.
1049 Lieber MR, Ma Y, Pannicke U, Schwarz K. 2003. Mechanism and regulation of
1050 human non-homologous DNA end-joining. *Nat Rev Mol Cell Biol* **4**: 712-720.
1051 MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. 2014. The Database of
1052 Genomic Variants: a curated collection of structural variation in the human
1053 genome. *Nucleic Acids Research* **42**: D986-D992.
1054 Makova KD, Yang S, Chiaromonte F. 2004. Insertions and deletions are male biased
1055 too: a whole-genome analysis in rodents. *Genome Res* **14**: 567-573.
1056 Martin M. 2011. Cutadapt Removes Adapter Sequences From High-Throughput
1057 Sequencing Reads. *EMBnetjournal* **17**: 10-12.
1058 Mcrae JF Clayton S Fitzgerald TW Kaplanis J Prigmore E Rajan D Sifrim A Aitken S
1059 Akawi N Alvi M et al. 2017. Prevalence and architecture of de novo mutations
1060 in developmental disorders. *Nature* **542**: 433-+.
1061 McVean G. 2007. What drives recombination hotspots to repeat DNA in humans?
1062 *Philosophical Transactions of the Royal Society London Series B Biological*
1063 *Sciences* **365**: 1213-1218.
1064 Messer PW. 2009. Measuring the Rates of Spontaneous Mutation From Deep and
1065 Large-Scale Polymorphism Data. *Genetics* **182**: 1219-1232.
1066 Messer PW, Arndt PF. 2007. The majority of recent short DNA insertions in the
1067 human genome are tandem duplications. *Mol Biol Evol* **24**: 1190-1197.
1068 Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer
1069 D, Bhandari A et al. 2012. Whole-genome sequencing in autism identifies hot
1070 spots for de novo germline mutation. *Cell* **151**: 1431-1442.
1071 Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G,
1072 Howie B, Karczewski KJ, Smith KS et al. 2013. The origin, evolution, and
1073 functional impact of short insertion-deletion variants identified in 179 human
1074 genomes. *Genome Res* **23**: 749-761.
1075 Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of
1076 recombination rates and hotspots across the human genome. *Science* **310**: 321-
1077 324.

1078 Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G,
1079 Donnelly P. 2010. Drive against hotspot motifs in primates implicates the
1080 PRDM9 gene in meiotic recombination. *Science* **327**: 876-879.
1081 Myers S, Freeman C, Auton A, Donnelly P, McVean G. 2008. A common sequence
1082 motif associated with recombination hot spots and genome instability in
1083 humans. *Nat Genet* **40**: 1124-1129.
1084 Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E,
1085 Maurano MT, Vierstra J, Thomas S et al. 2012. BEDOPS: high-performance
1086 genomic feature operations. *Bioinformatics* **28**: 1919-1920.
1087 Palmer N, Talib SZA, Ratnacaram CK, Low D, Bisteau X, Lee JHS, Pfeiffenberger E,
1088 Wollmann H, Tan JHL, Wee S et al. 2019. CDK2 regulates the NRF1/Ehmt1
1089 axis during meiotic prophase I. *J Cell Biol* **218**: 2896-2918.
1090 Powers NR, Parvanov ED, Baker CL, Walker M, Petkov PM, Paigen K. 2016. The
1091 Meiotic Recombination Activator PRDM9 Trimethylates Both H3K36 and
1092 H3K4 at Recombination Hotspots In Vivo. *PLoS Genet* **12**: e1006146.
1093 Pratto F, Brick K, Khil P, Smagulova F, Petukhova GV, Camerini-Otero RD. 2014.
1094 DNA recombination. Recombination initiation maps of individual human
1095 genomes. *Science* **346**: 1256442.
1096 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing
1097 genomic features. *Bioinformatics* **26**: 841-842.
1098 Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S,
1099 Dundar F, Manke T. 2016. deepTools2: a next generation web server for deep-
1100 sequencing data analysis. *Nucleic Acids Res* **44**: W160-165.
1101 Ramu A, Noordam MJ, Schwartz RS, Wuster A, Hurles ME, Cartwright RA, Conrad
1102 DF. 2013. DeNovoGear: de novo indel and point mutation discovery and
1103 phasing. *Nat Methods* **10**: 985-987.
1104 Reijns MAM, Kemp H, Ding J, de Proce SM, Jackson AP, Taylor MS. 2015.
1105 Lagging-strand replication shapes the mutational landscape of the genome.
1106 *Nature* **518**: 502-506.
1107 RK CY, Merico D, Bookman M, J LH, Thiruvahindrapuram B, Patel RV, Whitney J,
1108 Deflaux N, Bingham J, Wang Z et al. 2017. Whole genome sequencing
1109 resource identifies 18 new candidate genes for autism spectrum disorder. *Nat*
1110 *Neurosci* **20**: 602-611.
1111 Rodgers K, McVey M. 2016. Error-Prone Repair of DNA Double-Strand Breaks. *J*
1112 *Cell Physiol* **231**: 15-24.
1113 Sabarinathan R, Mularoni L, Deu-Pons J, Gonzalez-Perez A, Lopez-Bigas N. 2016.
1114 Nucleotide excision repair is impaired by binding of transcription factors to
1115 DNA. *Nature* **532**: 264-267.
1116 Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. 2004. JASPAR:
1117 an open-access database for eukaryotic transcription factor binding profiles.
1118 *Nucleic Acids Research* **32**: D91-D94.
1119 Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun
1120 V, Jaakkola T, Gifford DK. 2014. Discovery of directional and nondirectional
1121 pioneer transcription factors by modeling DNase profile magnitude and shape.
1122 *Nat Biotechnol* **32**: 171-+.
1123 Short PJ, McRae JF, Gallone G, Sifrim A, Won H, Geschwind DH, Wright CF, Firth
1124 HV, FitzPatrick DR, Barrett JC et al. 2018. De novo mutations in regulatory
1125 elements in neurodevelopmental disorders. *Nature* **555**: 611-616.

Sohni A, Tan K, Song HW, Burow D, de Rooij DG, Laurent L, Hsieh TC, Rabah R, Hammoud SS, Vicini E et al. 2019. The Neonatal and Adult Human Testis Defined at the Single-Cell Level. *Cell Rep* **26**: 1501-1517 e1504.

Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009. Human mutation rate associated with DNA replication timing. *Nat Genet* **41**: 393-395.

The 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68-74.

Turner TN, Coe BP, Dickel DE, Hoekzema K, Nelson BJ, Zody MC, Kronenberg ZN, Hormozdiari F, Raja A, Pennacchio LA et al. 2017. Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell* **171**: 710-722 e712.

Turner TN, Hormozdiari F, Duyzend MH, McClymont SA, Hook PW, Iossifov I, Raja A, Baker C, Hoekzema K, Stessman HA et al. 2016. Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. *Am J Hum Genet* **98**: 58-74.

van Gent DC, Hoeijmakers JH, Kanaar R. 2001. Chromosomal stability and the DNA double-stranded break connection. *Nat Rev Genet* **2**: 196-206.

Wang J, Tang C, Wang Q, Su J, Ni T, Yang W, Wang Y, Chen W, Liu X, Wang S et al. 2017. NRF1 coordinates with DNA methylation to regulate spermatogenesis. *FASEB J* **31**: 4959-4970.

Wellcome Trust Case Control C Craddock N Hurles ME Cardin N Pearson RD Plagnol V Robson S Vukcevic D Barnes C Conrad DF et al. 2010. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**: 713-720.

Werling DM, Brand H, An JY, Stone MR, Zhu L, Glessner JT, Collins RL, Dong S, Layer RM, Markenscoff-Papadimitriou E et al. 2018. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat Genet* **50**: 727-736.

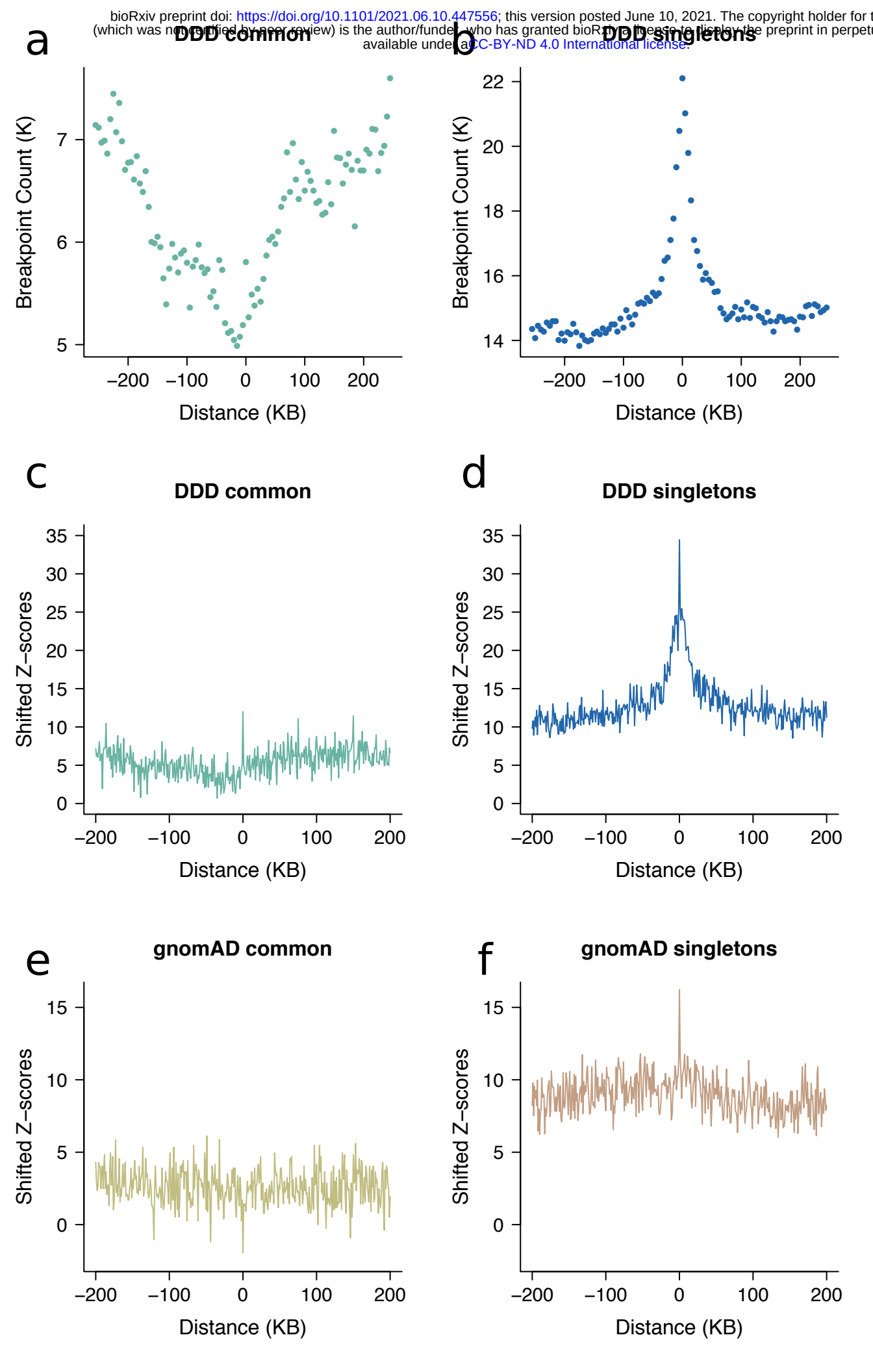
Wolfe KH, Sharp PM, Li WH. 1989. Mutation-Rates Differ among Regions of the Mammalian Genome. *Nature* **337**: 283-285.

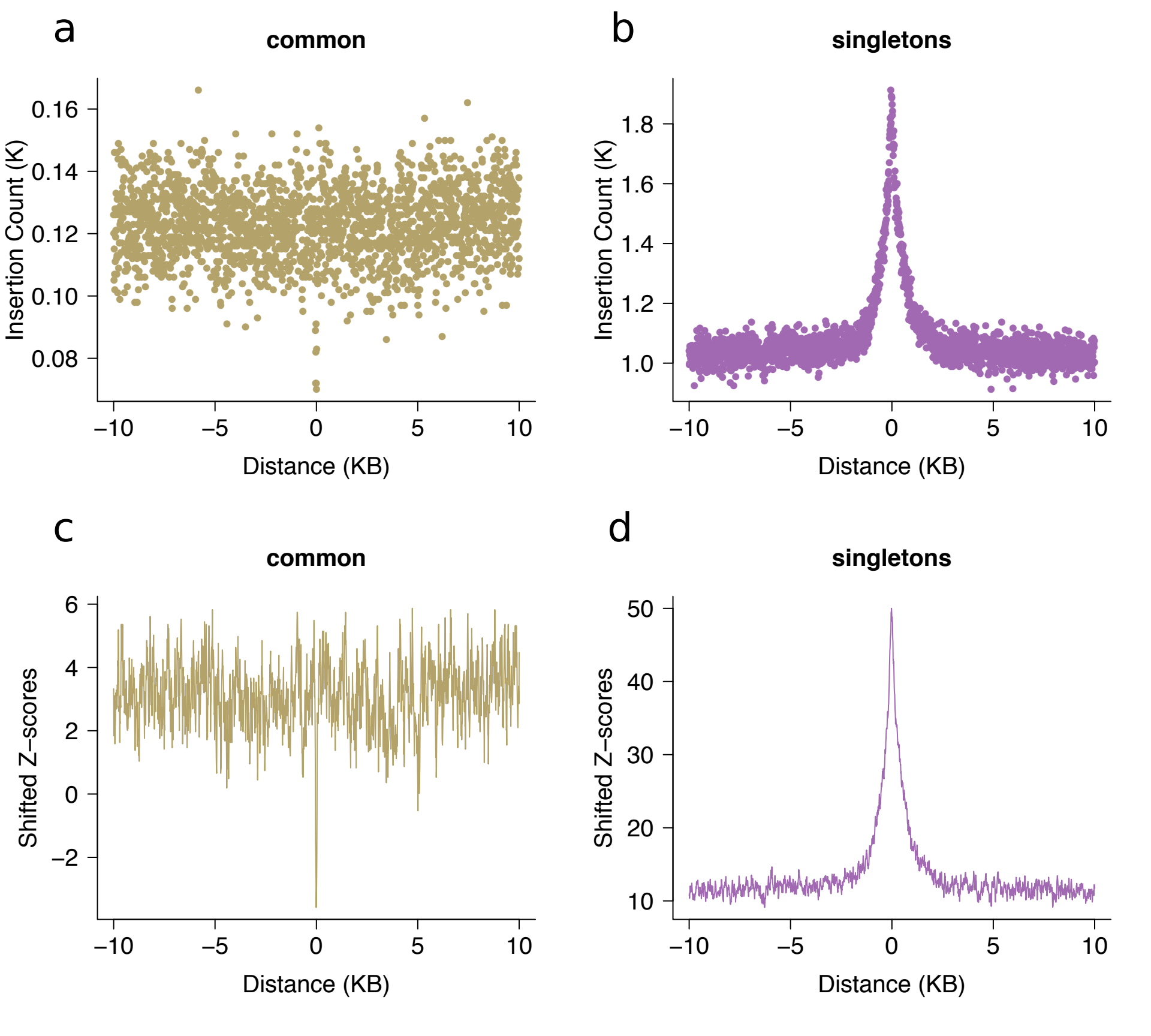
Yan F, Powell DR, Curtis DJ, Wong NC. 2020. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol* **21**: 22.

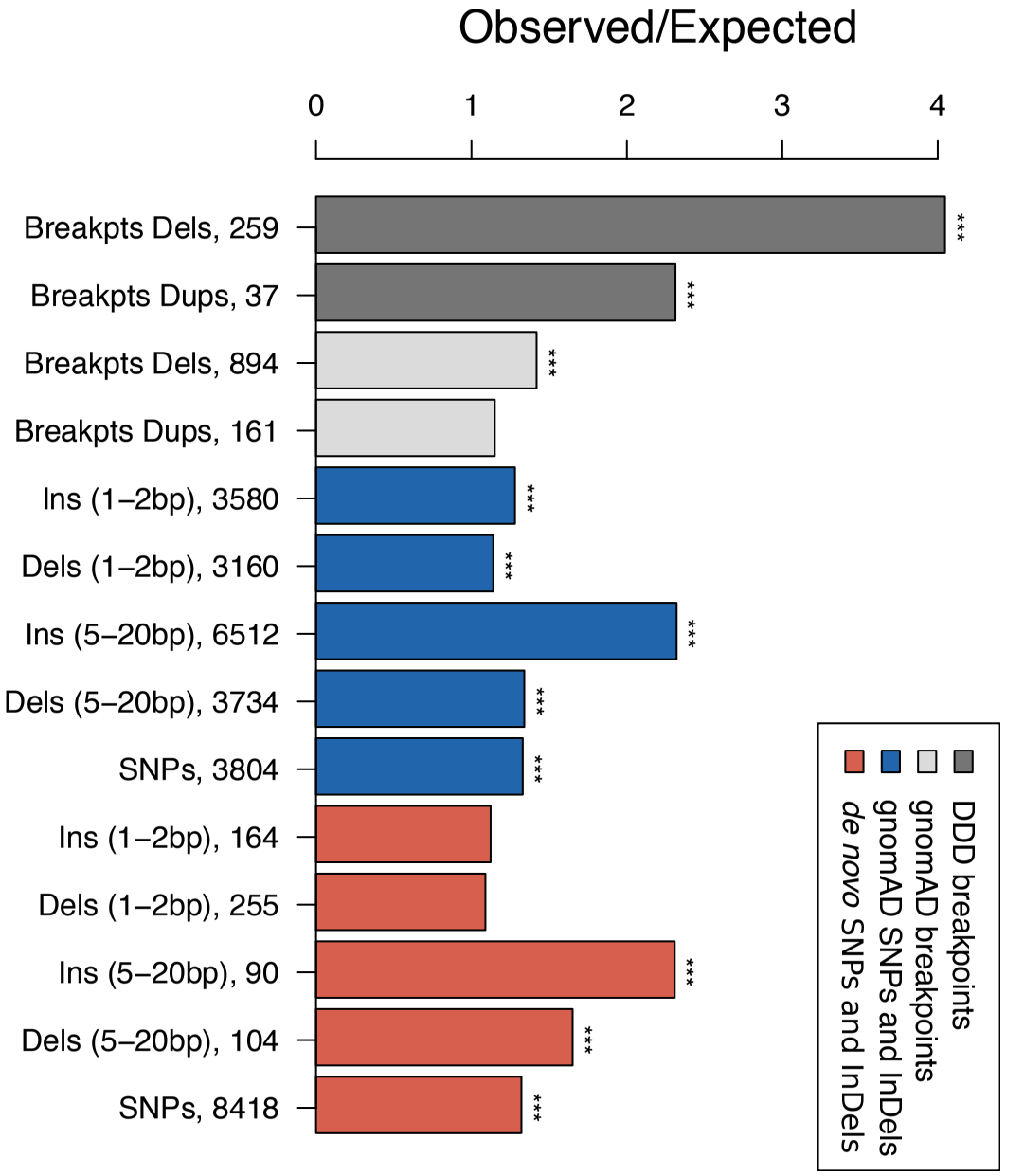
Yu G, Wang LG, He QY. 2015. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**: 2382-2383.

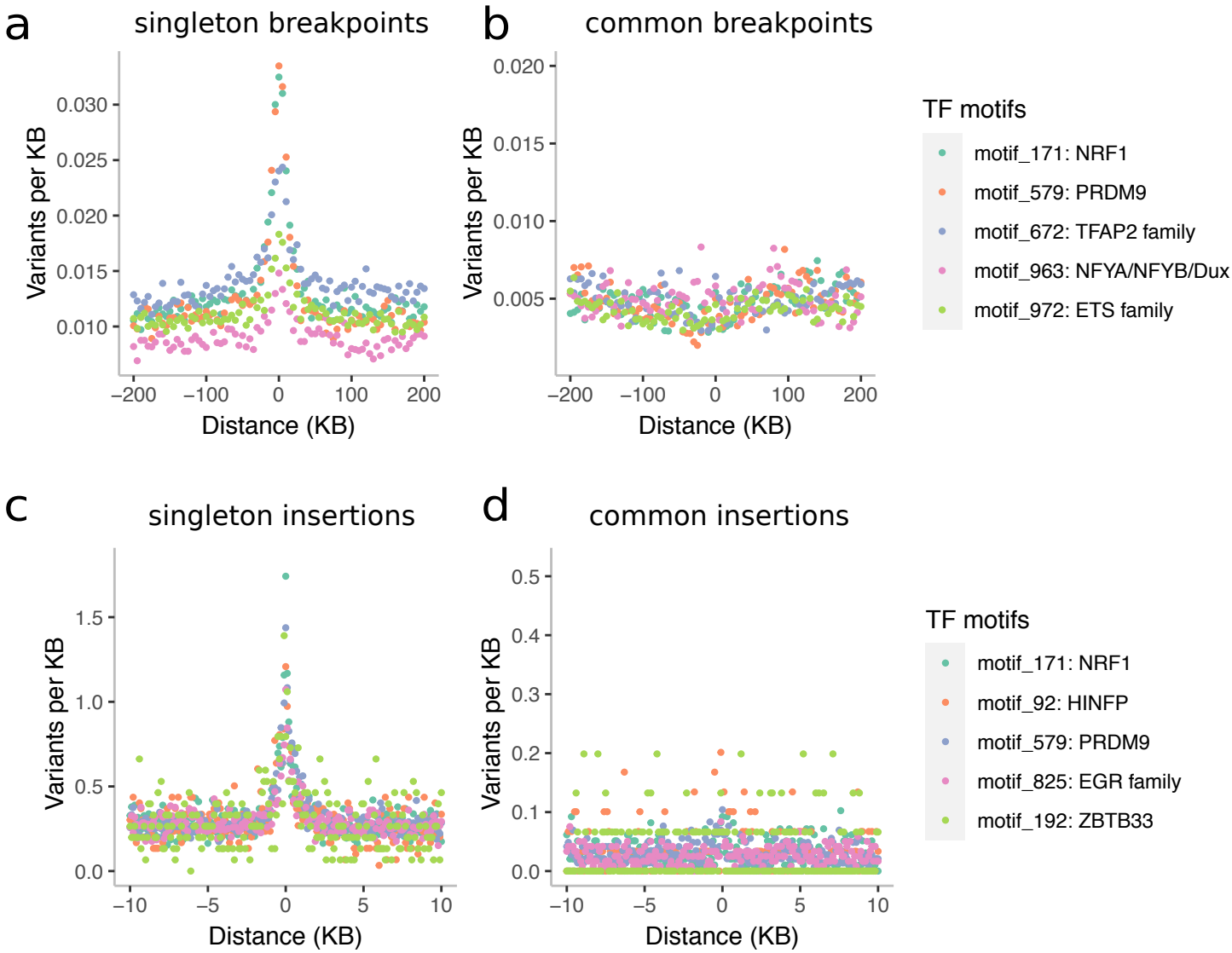
Yuen RK, Merico D, Cao H, Pellecchia G, Alipanahi B, Thiruvahindrapuram B, Tong X, Sun Y, Cao D, Zhang T et al. 2016. Genome-wide characteristics of de novo mutations in autism. *NPJ Genom Med* **1**: 160271-1602710.

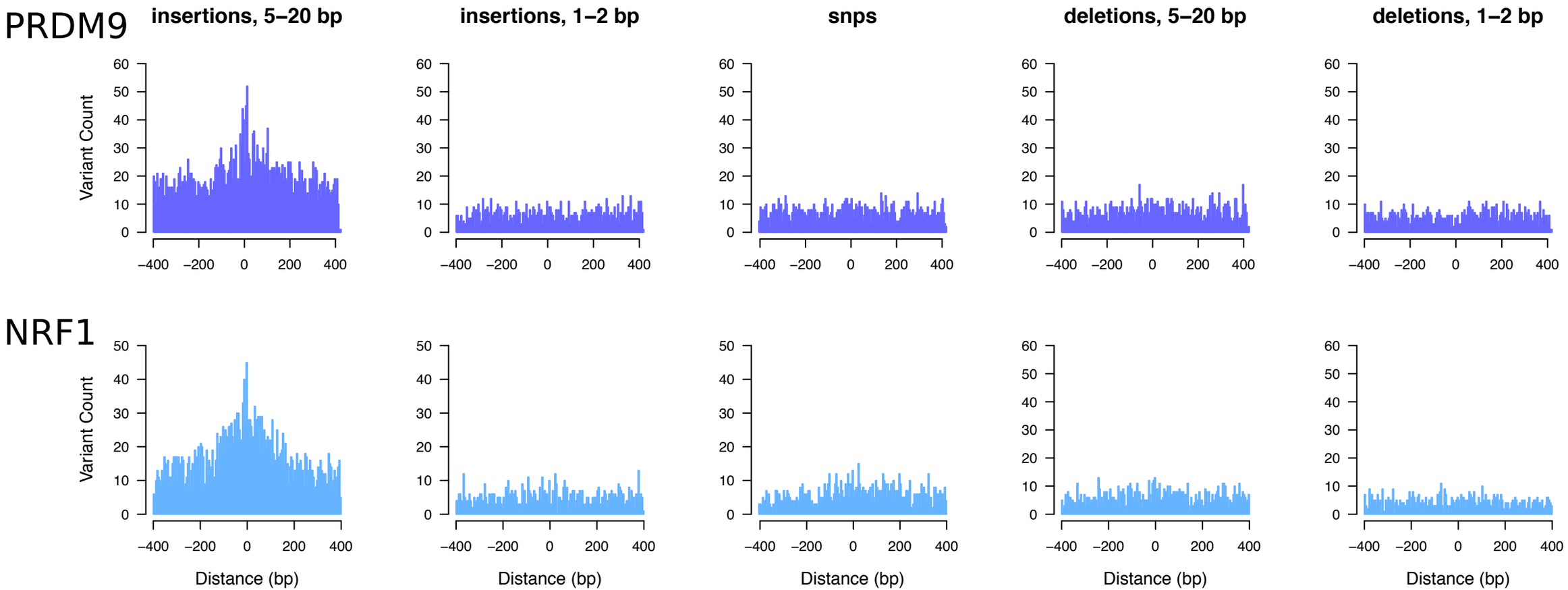
Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**.



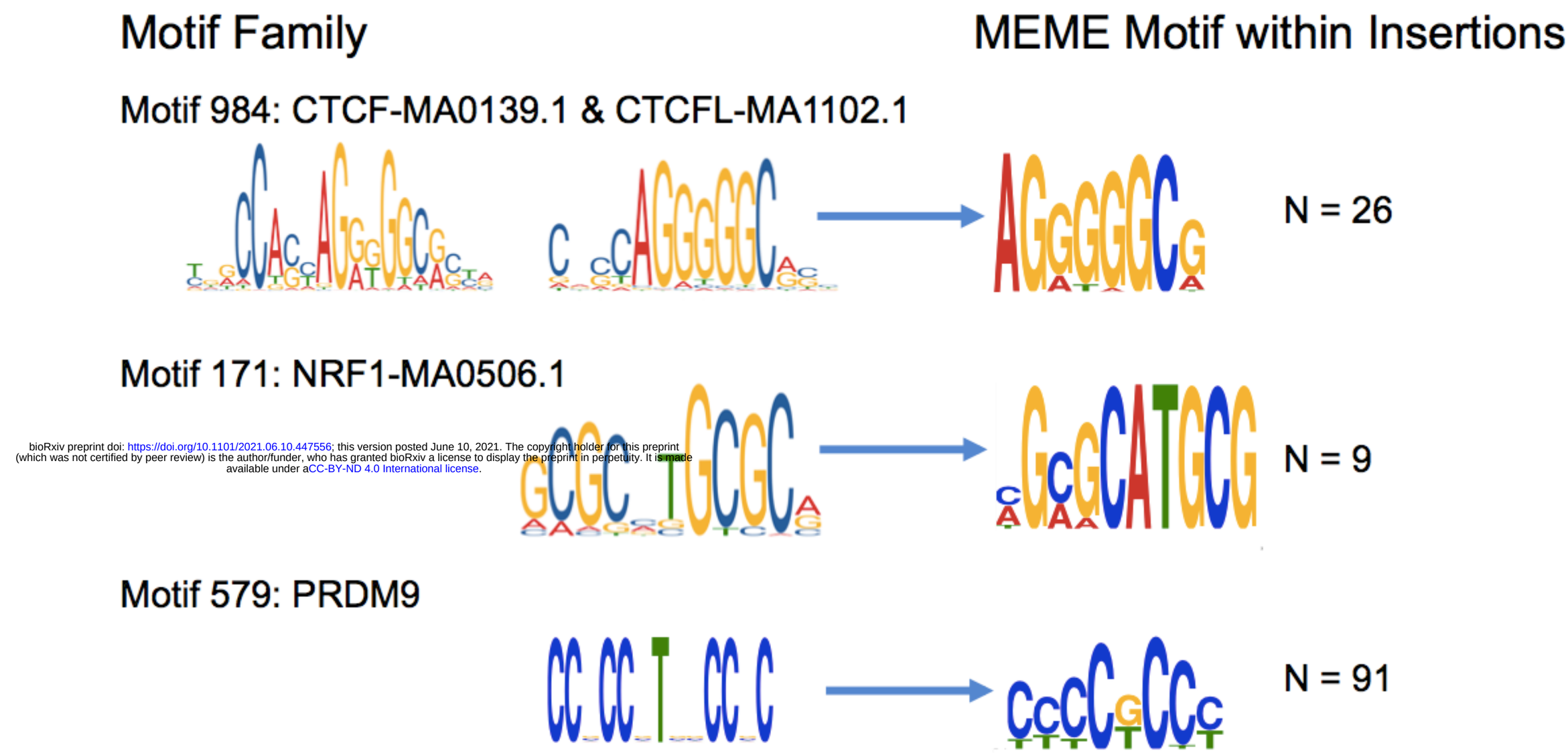

a
common
Insertion Count (K)
0.16
0.14
0.12
0.10
0.08
−10
−5
0
5
10
Distance (KB)
b
singletons
Insertion Count (K)
1.8
1.6
1.4
1.2
1.0
−10
−5
0
5
10
Distance (KB)
c
common
Shifted Z−scores
6
4
2
0
−2
−10
−5
0
5
10
Distance (KB)
d
singletons
Shifted Z−scores
50
40
30
20
10
−10
−5
0
5
10
Distance (KB)


Observed/Expected
0
1
2
3
4
Breakpts Dels, 259
Breakpts Dups, 37
Breakpts Dels, 894
Breakpts Dups, 161
Ins (1–2bp), 3580
Dels (1–2bp), 3160
Ins (5–20bp), 6512
Dels (5–20bp), 3734
SNPs, 3804
Ins (1–2bp), 164
Dels (1–2bp), 255
Ins (5–20bp), 90
Dels (5–20bp), 104
SNPs, 8418
***
DDD breakpoints
gnomAD breakpoints
gnomAD SNPs and InDels
de novo SNPs and InDels

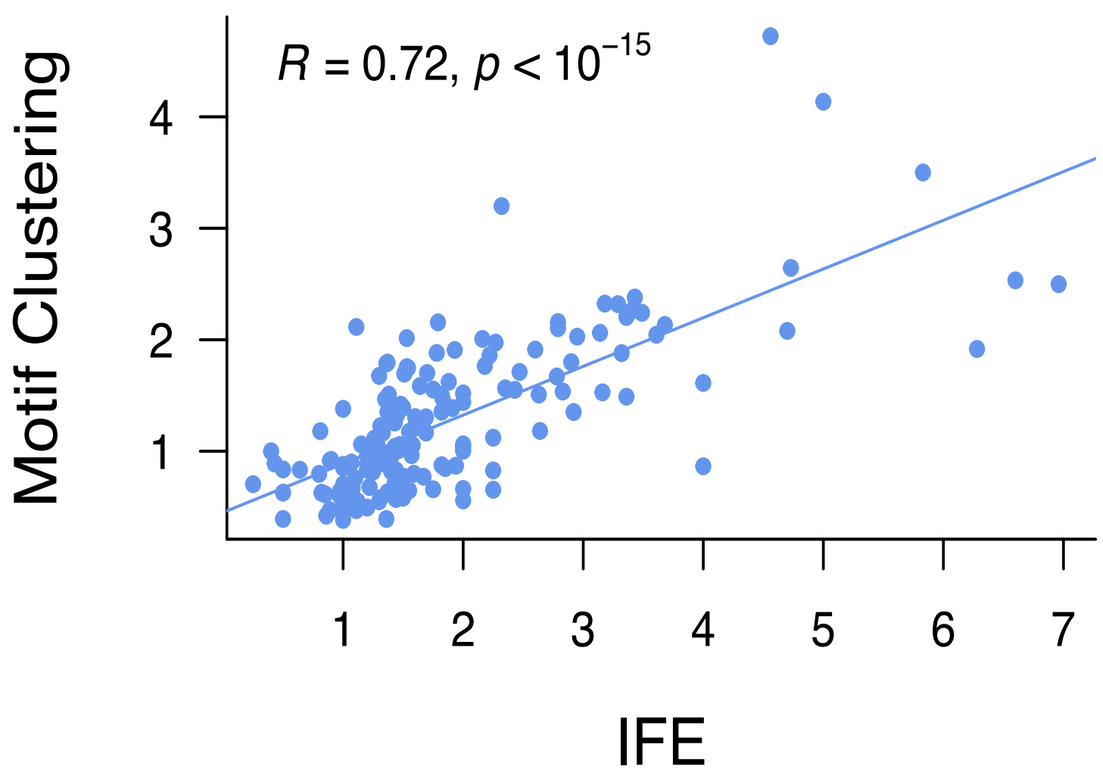

a
singleton breakpoints
b
common breakpoints
Variants per KB
Distance (KB)
TF motifs
motif_171: NRF1
motif_579: PRDM9
motif_672: TFAP2 family
motif_963: NFYA/NFYB/Dux
motif_972: ETS family
c
singleton insertions
d
common insertions
motif_92: HINFP
motif_825: EGR family
motif_192: ZBTB33


PRDM9
insertions, 5–20 bp
insertions, 1–2 bp
snps
deletions, 5–20 bp
deletions, 1–2 bp
NRF1
Variant Count
Distance (bp)
0
10
20
30
40
50
60
−400
−200
0
200
400

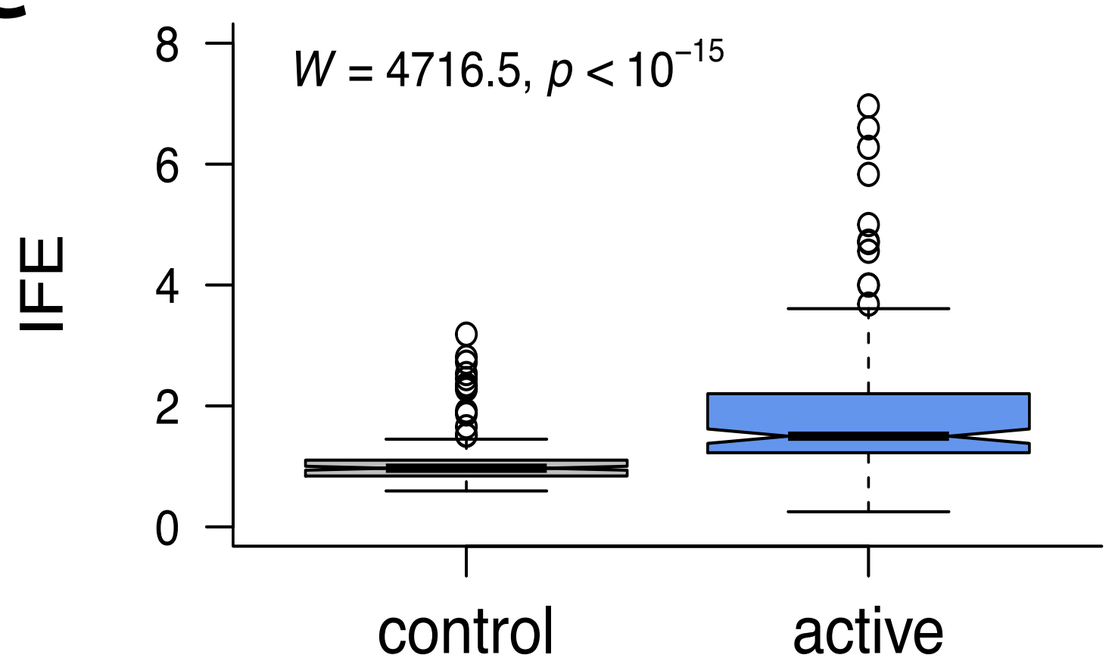
a

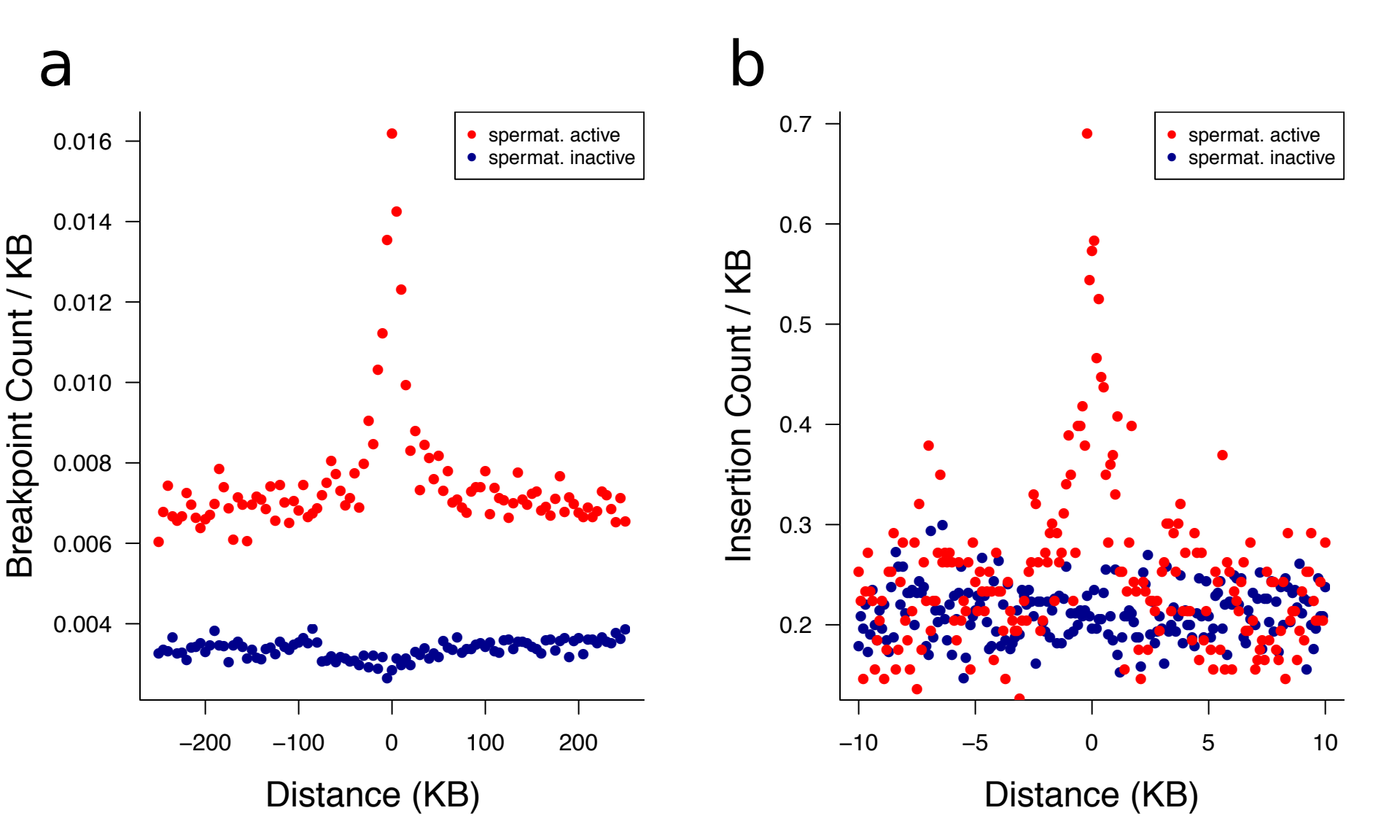


b



c




a
b
spermat. active
spermat. inactive
Breakpoint Count / KB
Insertion Count / KB
Distance (KB)
Distance (KB)
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.7
0.6
0.5
0.4
0.3
0.2
−200
−100
0
100
200
−10
−5
0
5
10