



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

Vera B. Kaiser, Lana Talmame, Yatendra Kumar, et al.

*Genome Res.* 2021 31: 1994-2007 originally published online August 20, 2021

Access the most recent version at doi:[10.1101/gr.275407.121](https://doi.org/10.1101/gr.275407.121)

---

**References** This article cites 93 articles, 11 of which can be accessed free at:  
<http://genome.cshlp.org/content/31/11/1994.full.html#ref-list-1>

**Open Access** Freely available online through the *Genome Research* Open Access option.

**Creative Commons License** This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

**Email Alerting Service** Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).



---

To subscribe to *Genome Research* go to:  
<https://genome.cshlp.org/subscriptions>

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

Vera B. Kaiser,<sup>1</sup> Lana Talmane,<sup>1</sup> Yatendra Kumar,<sup>1</sup> Fiona Semple,<sup>1</sup> Marie MacLennan,<sup>1</sup> Deciphering Developmental Disorders Study,<sup>1,2</sup> David R. FitzPatrick,<sup>1</sup> Martin S. Taylor,<sup>1,3</sup> and Colin A. Semple<sup>1,3</sup>

<sup>1</sup>MRC Human Genetics Unit, MRC Institute of Genetics and Cancer, The University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, United Kingdom; <sup>2</sup>The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom

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 NRFI, are also associated with increased rates of SV breakpoints, apparently independent of recombination. Singleton short insertions ( $\geq 5$  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.

[Supplemental material is available for this article.]

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 multimegabase ( $>10^6$  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, whereas CNVs arising from nonhomologous mechanisms are enriched in late replicating regions (Koren et al. 2012). Local chromatin structure also influences the mutation rate. However, finer-scale variation ( $<1$  Mb) 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, 2018). Transcription factor binding sites (TFBSs) are particularly prone to point mutations in cancer (Kaiser et al. 2016), probably owing 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), as well as population genetic-based measures of recombination hotspots (HSS) (Myers et al. 2005; The 1000 Genomes Project Consortium 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 focusing on extremely rare variants (Messer 2009; Carlson et al. 2018; Li and Luscombe 2020). Making use of such variant data sets as well as novel ATAC-seq data in spermatogonia, we study the mutational landscape at TFBSs in accessible human spermatogonial chromatin.

<sup>3</sup>These authors contributed equally to this work.

Corresponding author: [vera.kaiser@ed.ac.uk](mailto:vera.kaiser@ed.ac.uk)

Article published online before print. Article, supplemental material, and publication date are at <https://www.genome.org/cgi/doi/10.1101/gr.275407.121>. Freely available online through the *Genome Research* Open Access option.

© 2021 Kaiser et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

## Results

### Spermatogonial regulatory regions are enriched for rare deletion breakpoints

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 premeiotic 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; Supplemental Datasets 1–3), and multiple metrics (Supplemental Fig. S1A–C) indicated high data quality (Yan et al. 2020). Hierarchical clustering (Ramírez et al. 2016) showed that this novel spermatogonial ATAC-seq data set displays a genome-wide distribution of peaks consistent with other spermatogonial-derived data and is distinct from ES cell and somatic tissue data sets (Supplemental Fig. S2).

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 ultrarare 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 (An et al. 2018; <http://denovo-db.gs.washington.edu/>; <https://research.mss.ng/>). Based on the DDD data set, a combination of high-density arrayCGH and exome sequencing (Deciphering Developmental Disorders Study 2015), we identified 6704 singleton deletion variants among 9625 DDD probands (carrier frequency of ~0.002% in the combined data set) (Supplemental Table S1).

Permutation analysis shows that DDD singleton breakpoints are enriched at spermatogonial ATAC-seq sites, their overlap being more than four 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 (Fig. 1B,D). We also considered 6013 deletions (represented by their unique breakpoint coordinates; see Methods) that were present in the DDD consensus data set (Deciphering Developmental Disorders Study 2015) at a frequency of at least 1%, representing variants expected to be relatively common in human populations (Methods; Supplemental Table S1). These variants show a dip in frequency and downward trend near active sites (Fig. 1A,C). However, we note that the overlap between common variant breakpoints and ATAC-seq peaks is still approximately twofold 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 with neighboring 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 data set of human variants (Fig. 1F) from whole-genome sequence (WGS) analysis (Supplemental Table S1; Collins et al. 2020). We again find a significant enrichment of singleton variant breakpoints at ATAC-seq peaks, and this enrichment is not seen for common variants (Fig. 1E).

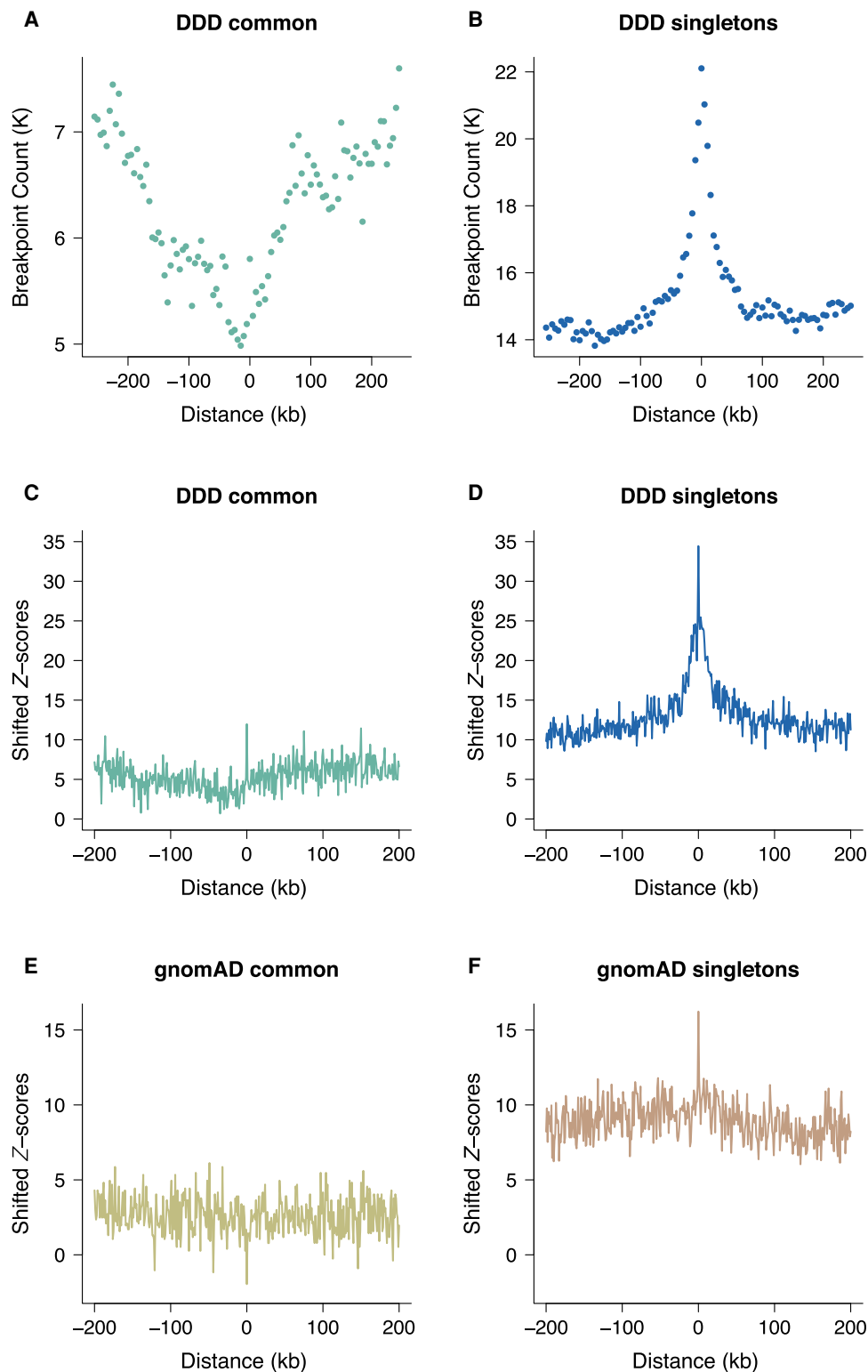
### Locally elevated mutation at spermatogonial TFBSs

Compared with 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 six new variants per genome and generation (Besenbacher et al. 2016). Short indels ( $\leq 4$  bp) are thought to arise owing 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  $\leq 20$  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 (Methods; Supplemental Tables S3, S4; Li et al. 2019). The rate of singleton 5- to 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 (Fig. 2B); shifted Z-scores based on genome-wide circular permutations similarly show a highly localized spike of insertions around TFBSs (Fig. 2D). This pattern starkly contrasts the localized depletion of common variants of the same mutation class at the same binding sites (Fig. 2A,C), 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 (Fig. 3), and in the gnomAD data set, in which all singleton classes have been ascertained by WGS, the enrichment is strongest for insertions  $\geq 5$  bp. We confirmed the enrichment of singleton short insertions and SV deletion breakpoints at spermatogonial TFBSs, using an independent permutation approach with BEDTools' "bedtools shuffle" (Supplemental Methods; Supplemental Table S5; Quinlan and Hall 2010).

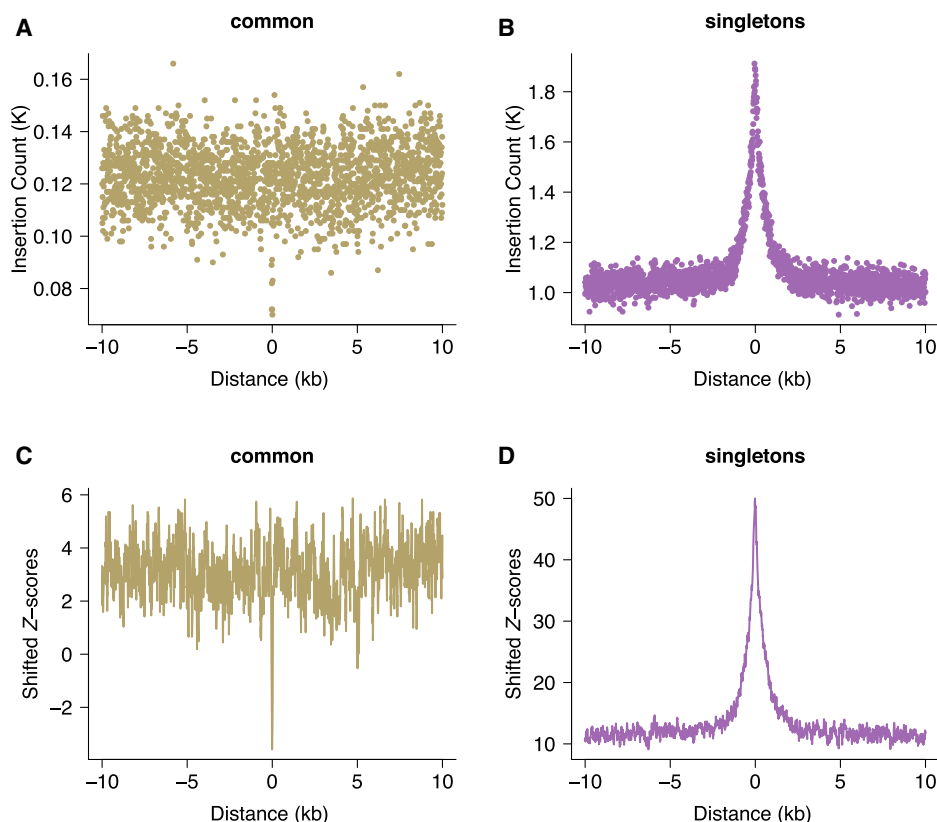
In addition to singleton variants from large population samples, we also compiled a set of gold standard de novo short variants from a range of trio sequencing studies (see Methods). The de novo variants show a similar trend to the gnomAD singleton variants, with a moderate (~10%–60%) increase of mutation rates at TFBSs for all categories of short 1- to 2-bp sequence variants, but larger increase of ~130% for insertions of 5–20 bp (Fig. 3). These results were confirmed using a set of independent positive and negative control sites (Supplemental Fig. S3A,B). We conclude that regulatory sites that are active in spermatogonia show unusual parallel enrichments for both large SV breakpoints and 5- to 20-bp insertions, consistent with localized DNA damage or error-prone repair.

### Germline PRDM9 and NRFI binding generate HSs for structural variation

To examine any differences in mutational loads associated with different binding factors, we analyzed 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



**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, centered around the midpoints of spermatogonial ATAC-seq peaks. “Singletons” are breakpoints of deletions with a frequency of  $\sim 0.002\%$  across population samples; “common” variants are seen at a frequency of at least 1% in the DDD consensus data set (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, centered around spermatogonial ATAC-seq peaks. “Singletons” are breakpoints of deletions with a frequency of  $\sim 0.002\%$  across population samples; “common” variants are seen at a frequency of at least 5% in the gnomAD V.2 data set. 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 focused on spermatogonial binding sites. Insertion count (A,B) and shifted Z-scores (C,D) of gnomAD singleton and common insertions (5–20 bp), centered around spermatogonial TFBSs. Singletons are seen only once in the gnomAD V.3 data set (allele frequency  $\leq 0.001\%$ ) 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 ( $P < 10^{-4}$ ).

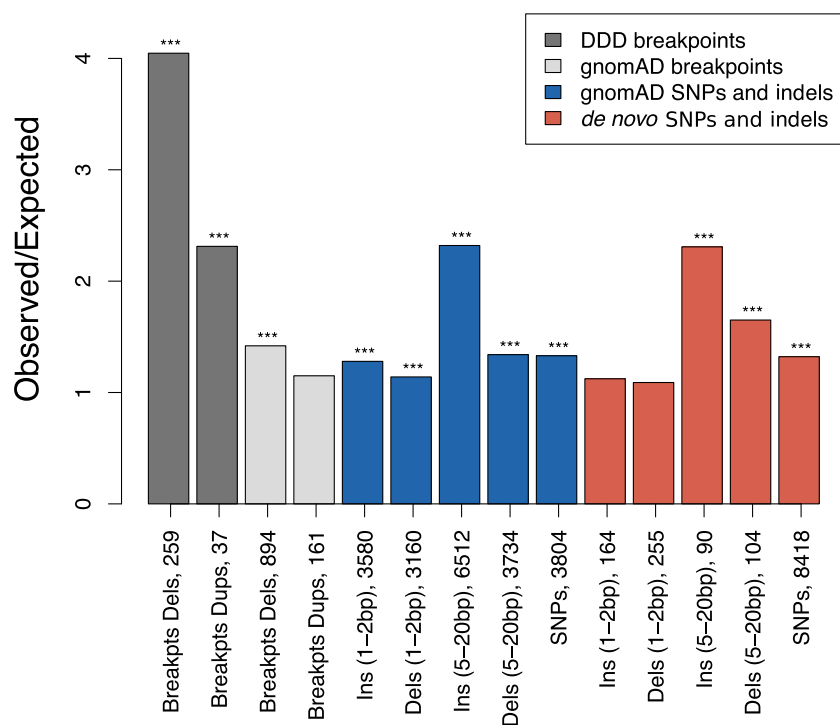
families (Supplemental Table S6). Furthermore, using the reported binding site motif for PRDM9 (Myers et al. 2008), we defined 9778 putative PRDM9-bound sites corroborated by evidence for H3K4me3 enrichment in testis (Methods).

The spermatogonial binding sites of 11% (19/167) of motif families overlapped DDD singleton deletion breakpoints more often than expected, and similarly, 29% (48/167) of motif families were significantly enriched for gnomAD singleton deletion breakpoints (Bonferroni-corrected  $P = 0.017$ ); no motif family was found to be depleted for breakpoints in either data set (Supplemental Tables S3, S4), suggesting that increased load is a common feature of TFBSs bound by different transcription factors in the germline. Similarly, singleton 5- to 20-bp 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), namely, 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 8139 PRDM9 sites with a distance of at least 500 bp from HSs and ssDNA sites, respectively). PRDM9 sites are similarly associated with higher rates of singleton deletion breakpoints (Fig. 4A,C), 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 nuclear respiratory factor 1 (NRF1; IFE=7.0) and HINFP (IFE=6.6) exceed the disruption seen at PRDM9 sites, and NRF1 sites are also disrupted at high rates according to DDD and gnomAD breakpoint data (Supplemental Tables S3, S4). Shifted Z-scores for the enrichment of short insertions (5–20 bp) at both the 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 S6, S7). 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 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, 2018). Reprocessing these previous data sets identically to our own reveals that PRDM9, NRF1, and HINFP sites are again among the top five disrupted motif families (Supplemental Tables S8, S9).

Although both PRDM9 and NRF1 binding sites are GC rich, their modest motif similarity suggests that the two factors occupy



**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: (\*\*\*)  $P < 0.001$ . The type of variant tested and the total number of observed variants overlapping spermatogonial TFBSs are indicated below each bar.

distinct binding motifs (PWMclus: Pearson's correlation distance  $r = 0.35$  for PRDM9 vs. 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 (1199) ATAC-seq peaks contained both NRF1 and PRDM9 binding motifs. The disruption of motifs within these cobound 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, 1311 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 with all CTCF motifs (Supplemental Table S4), as was PRDM9 (ratio=8.2; observed: 115; expected: 14;  $P < 10^{-4}$  in each case).

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 (fewer than 10,000 sites each) (Supplemental Tables S3, S4).

In general, mutational loads appear to be dependent on the level of chromatin accessibility (MACS2 peak scores) (Zhang et al. 2008) 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 four factors incur the highest indel loads (Supplemental Fig. S4A–D). The significant positive correlation between the rates of

binding site disruption via singleton insertions and deletion breakpoints across all motif families (Spearman's  $R = 0.52$ ;  $P < 10^{-5}$ ) (Supplemental Fig. S5A–C) 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 owing to erroneous variant calling in the singleton data set because we observe similar patterns for common variants of the same variant categories (genome-wide  $Z$ -score = 62.9;  $P < 10^{-4}$ ).

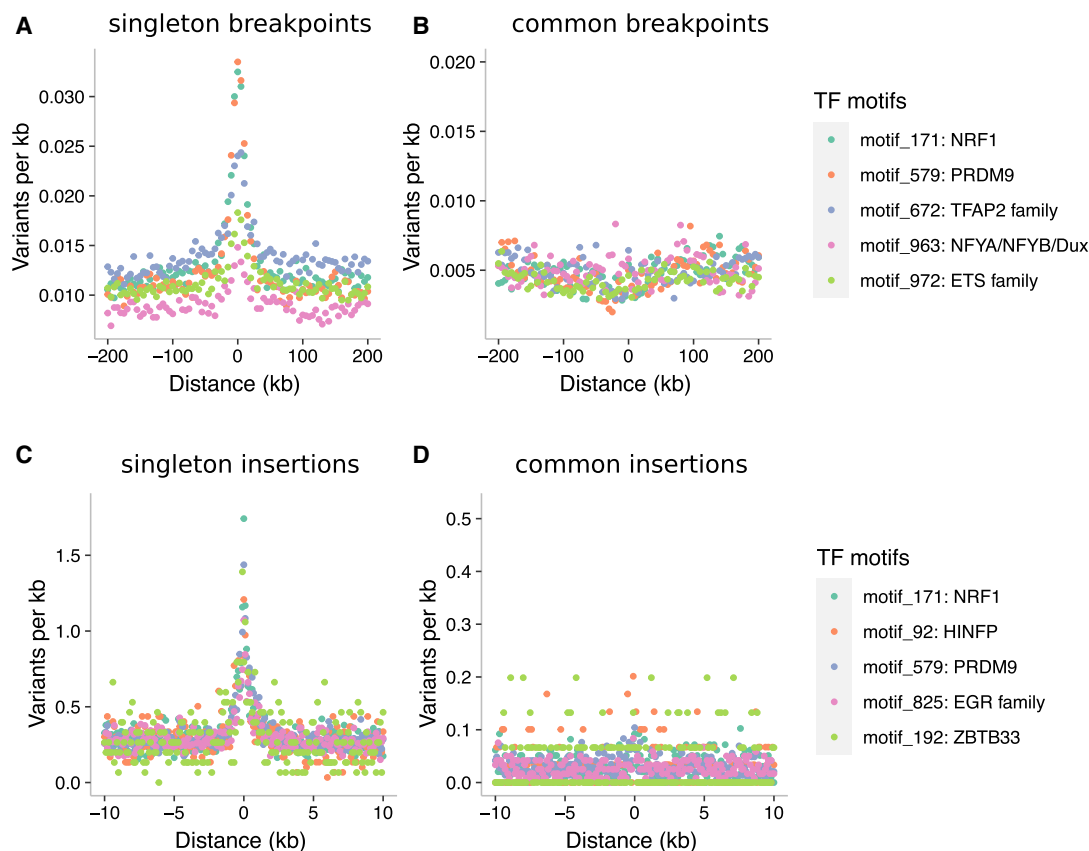
### Short insertions generate clustered binding sites within regulatory regions

Five- to 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 (Fig. 5; Supplemental Fig. S7). Despite a moderate genome-wide enrichment (Fig. 3), the 1- to 2-bp insertions characteristic of polymerase slippage do not peak in the immediate neighborhood of TFBSs (Fig. 5; Supplemental Fig.

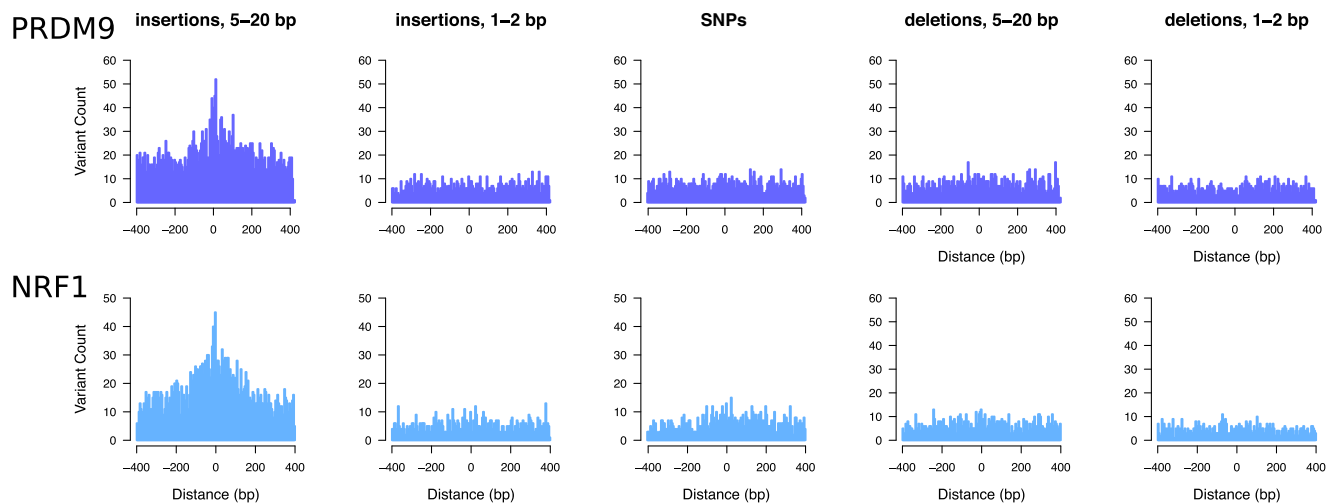
S7). We examined the consequences of elevated 5- to 20-bp insertion rates at TFBSs using an exhaustive motif search algorithm (Bailey et al. 2009), which finds overrepresented sequence motifs among a set of input sequences. We found that the inserted sequences at a mutated TFBS often contain additional copies of the sequence motif corresponding to the original TFBS (Fig. 6A; 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 data set (ranked 12th out of 167 motif families, based on the number of insertions per TFBS) (Supplemental Table S4). We find that spermatogonial active sites show a greater enrichment of singleton insertions than control binding sites outside ATAC-seq peaks (Fig. 6C). Combined with a positive correlation between homotypic motif clustering and insertion rate (Fig. 6B), this suggests that spermatogonial 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 50-bp regions flanking singleton insertion and deletion breakpoints, we were able to rediscover the sequence motifs of commonly disrupted binding sites, including the motifs of PRDM9 and NRF1 (Supplemental Table S10). In contrast, genome-wide, the motifs discovered flanking these variants were

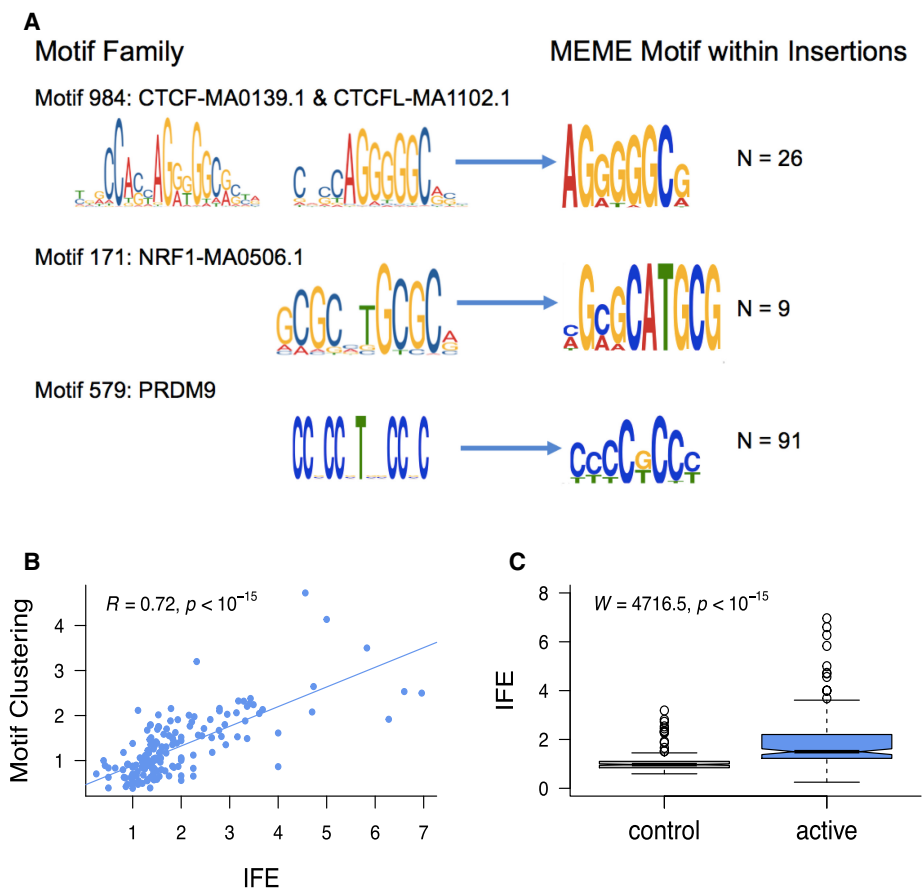
## Elevated mutation rates at spermatogonial TFBSs



**Figure 4.** Binding factors associated with the highest rates of mutation at spermatogonial binding sites. Plots are centered on the binding sites of a given motif family inside ATAC-seq footprints. (A,B) Singleton (A) and common (B) deletion breakpoints in the DDD cohort; singletons are breakpoints of deletions with a frequency of  $\sim 0.002\%$  across population samples; common variants are seen at a frequency of at least 1% in the DDD consensus data set. (C,D) Singleton (C) and common (D) insertions (5–20 bp) in the gnomAD data set. Singletons are seen only once in gnomAD V.3 (allele frequency  $\leq 0.001\%$ ), and common variants have an allele frequency of at least 5% within gnomAD V.3. Only 10-kb regions around TFBSs with  $\geq 95\%$  unique mappability (umap24 scores) were included. The top five 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  $\geq 95\%$  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 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.

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.

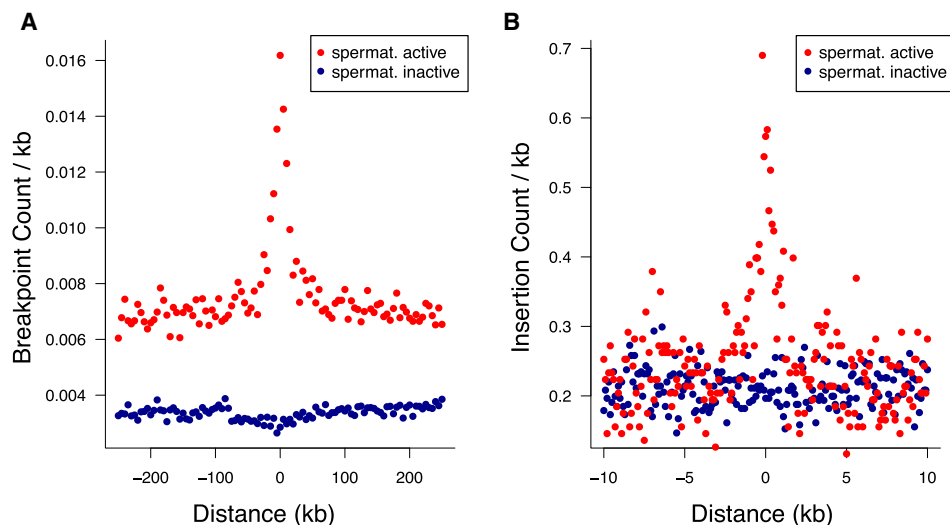
### Genomic instability at spermatogonial TFBSs impacts enhancers active in neural development

Because 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 (26,162 brain enhancers) in the male germline. We then calculated the odds ratio of a singleton mutation affecting a brain enhancer that is also active in spermatogonia relative to a brain enhancer that 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 (Fig. 7A,B). 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  $< 10^{-15}$ ), suggesting a link between enhancer activity, the sharing of enhancers across tissues, and propensity to mutation. The subset of brain enhancers that 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



**Figure 7.** Neural enhancers with activity in spermatogonia suffer elevated mutation rates. (A,B) Singleton DDD deletion breakpoint (A) and singleton gnomAD insertion (B; 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  $\geq 95\%$  unique mappability (umap24 scores) were included (3409 brain enhancers that are inactive in spermatogonia and 1029 that are active).

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 (Fig. 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 S11). In addition, control motif sites (representing the same TFBS but located outside of ATAC-seq peaks) are subject to lower rates of mutation compared with motifs within spermatogonial ATAC-seq peaks (Fig. 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 because some control sites may be bound at time points in the germline that our ATAC-seq data cannot ascertain.

Because 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 S12). 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 with genome-wide expectations, suggesting that the inferred effects of protein binding on mutation are larger than the reduction in mutation owing 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 HSs, 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 data set (only regions with high mappability were included in our more detailed analyses of TFBSs) (Figs. 3–7; 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- to 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 shown 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- to 20-bp insertions. 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. These

results have implications for the evolution of binding site patterns within regulatory regions, as well as 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 compartmentalization 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 may originate from DNA breakage, followed by misrepair, conceivably via a pathway such as nonallelic 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 2010; Myers et al. 2008; Berg et al. 2010). For example, the locations of PRDM9 binding HSs coincide with recurrent SV breakpoints causing Charcot-Marie-Tooth disease, as well as Hunter and Potocki-Lupski/Smith-Magenis syndromes (Pratto et al. 2014). It is possible that the sequence similarity at TFBSs scattered across the genome may make them particularly prone to NAHR. However, the sequence similarity between the low copy repeat units, known to be involved in NAHR, is usually of the size of several kilobases (Gu et al. 2008), rather than sequences on the scale of TFBSs. The NHEJ pathway can also lead to short insertions after DNA breakage, usually in G<sub>0</sub> and G<sub>1</sub> phases of the cell cycle. Indeed, NHEJ is the most common repair pathway of DSBs in mammals, and it is typically error prone (van Gent et al. 2001; Lieber et al. 2003). During NHEJ, DSB ends are resected to form single-stranded overhangs, but when pairing occurs between the tips of the overhangs, sequences near the breakpoints will often be duplicated (Rodgers and McVey 2016). Two previous studies using human-chimpanzee-macaque multiple alignments have shown that high numbers of short insertions have occurred in the human lineage (Kvikstad et al. 2007; Messer and Arndt 2007), and both conclude that these insertions preferentially take place in the male germline, evidenced by decreased mutation rates on the X Chromosome, with similar observations in rodents (Makova et al. 2004).

The data presented here suggest that different DNA-binding proteins differ widely in their impact on mutation rates. The two proteins with the largest impacts, NRF1 and PRDM9, are both highly expressed in testis, revealing a possible link between the expression level of a gene encoding a DNA-binding protein and the propensity for breakage or inefficient repair at the sites the protein binds. Incidentally, *NRF1* has a 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 The 1000 Genomes Project Consortium 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, suggest 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 suggest a potential solution to this paradox, in which 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 h 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 labeled with a 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 Biosciences) based on PE fluorescence and cell shape, according to forward/side scatter. Isolated cells were subjected to ATAC-seq using the protocol and reagents previously described (Buenrostro et al. 2013), followed by paired-end sequencing on Illumina HiSeq 4000 (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 was combined sequencing runs H.5.1 and H.5.4; a noncancer patient; large cells, high side scatter; and 58,000 and 42,000 cells, respectively. Replicate 2 was combined sequencing runs H.5.2 and H.5.5, the same noncancer patient as Replicate 1, large cells, and 36,000 and 23,000 cells, respectively. Replicate 3 was combined sequencing runs H.7.3 and H.10.2, normal tissue from cancer patients, large cells, and 69,000 and 24,000 cells, respectively. Raw reads were processed and ATAC-seq peaks called as described in the [Supplemental Methods](#). For the downstream mutation analyses, ATAC-seq peaks from Replicates 1 and 2 (the noncancer patient) were merged, creating a single peak set. This data set 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 matrices (PWMs) of the JASPAR 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 (because 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 S7). 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. Because 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 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 premeiotic: 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 data sets

Using the same procedure as described in the Supplemental Methods, we processed raw ATAC-seq reads from previously published data sets in order to call MACS2 peaks from short sequencing fragments (Zhang et al. 2008). Data sets included ATAC-seq reads from the germinal zone and cortical plate of the developing brain (NCBI Sequence Read Archive [SRA; <https://www.ncbi.nlm.nih.gov/sra>] accession numbers SRR6208926, SRR6208927, SRR6208938, SRR6208943) (de la Torre-Ubieta et al. 2018), ATAC-seq experiments of KIT<sup>+</sup> spermatogonia (SRA accessions SRR7905001 and SRR7905002) (Guo et al. 2018), SSEA4<sup>+</sup> spermatogonia (SRA: SRR5099531, SRR5099532, SRR5099533, SRR5099534) (Guo et al. 2017), and ESC cells (SRA: 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 data sets (liver,

ENCFF628MVCV; ovary, ENCFF780JBA; spleen, ENCFF294ZCT; testis, ENCFF048IOT; transverse colon, ENCFF377DAO) (The ENCODE Project Consortium 2012; Davis et al. 2018) were downloaded as BAM files and converted to BEDPE format, and short fragments were identified for peak calling.

### Structural variant breakpoint data

Large SVs, identified by high-density arrayCGH or a combination of arrayCGH + exome sequencing, were extracted from a cohort of 9625 DDD patients, using variant calling procedures as described previously (Deciphering Developmental Disorders Study 2015). We filtered the DDD variants to only keep variants that fulfilled the following criteria: a CNsolidate `wscore`  $\geq 0.468$ , a call  $P < 0.01$ , and a mean  $\log_2$  ratio of  $< -0.41$  for deletions and 0.36 for duplications; CIPHER “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 data sets (80% reciprocal overlap criterion). Because there are 9625 patients in the DDD data set, the `gnomAD V.2` data set contains SVs from 10,738 genomes, and the `dgv` contains SVs from 29,084 individuals, this puts an upper limit of the frequency of carriers of a singleton variant at  $\sim 0.002\%$ . Breakpoints were identified as the 5' and 3' coordinates of SVs, resulting in 13,406 singleton deletion and 3406 duplication breakpoints; the resolution of the breakpoints was such that the median and mean confidence intervals were 300 bp and 12 kb, respectively. Thus, the DDD data set has a lower resolution compared with the WGS data, but its advantage is that it does not suffer from mapping and variant calling issues associated with the latter (Mahmoud et al. 2019).

We further identified 11,962 “common” deletion variants in the DDD data set, which had a minimum variant frequency of 1% in the consensus CNV data set as described by the DDD study (2015), namely, pooled CNV data sets of Conrad et al. (2010), The 1000 Genomes Project Consortium (2010), the Wellcome Trust Case Control Consortium (2010), and the DDD normal controls. We used the 80% reciprocal overlap criterion and grouped common variants using the `bedmap` options `--echo-map --fraction-both 0.8`, followed by `bedops --merge` (Neph et al. 2012). The breakpoints of common variants are thus the outermost coordinates of all SVs that are collapsed into a given variant. The overlap of such “common” breakpoints with ATAC-seq peaks was assessed independently of SV allele frequencies; namely, a group of common SVs contributed two breakpoints to the analysis, and this number was further reduced if one breakpoint coordinate was shared between two common SVs, so as to only count each common breakpoint once.

We also identified a set of singleton CNVs called with the Manta algorithm (Chen et al. 2016) from the `gnomAD V.2` database (80% reciprocal overlap criterion with `gnomAD V.2`, `dgv` and DDD variants) (Collins et al. 2020), resulting in a set of 73,063 deletion and 15,419 duplication breakpoints seen in  $\sim 0.002\%$  of individuals but called with a different approach compared with the DDD. Common deletions and duplications ( $P \geq 0.05$ ) were also extracted from the `gnomAD V.2` data set; these variants had also been called with the Manta algorithm and included 5954 deletion and 1586 duplication breakpoint sites.

### 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) (Danecek et al. 2011). Multiallelic variants were split using BCFtools (Danecek et al. 2021), 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 or more; that is, the allele frequency of singletons was  $P \leq 0.001\%$ .

We subdivided gnomAD indels into singleton insertions and deletions of different sizes: 1–2 bp (most commonly arising owing to replication slippage) and those 5–20 bp (arising owing 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 WGS studies (Michaelson et al. 2012; Ramu et al. 2013; The Genome of the Netherlands Consortium 2014; Besenbacher et al. 2015; Turner et al. 2016, 2017; Yuen et al. 2016, 2017; Jonsson et al. 2017; Werling et al. 2018), which included a total of 404,238 variants from 4560 samples. Additional samples, which were not already included in the denovo-db data set, were downloaded from the MSSNG database (<https://research.mss.ng/>), version 2019/10/16, which added 2243 samples and 215,044 de novo mutations. A third source of de novo variants came from An et al. (2018): 3805 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 (Gel et al. 2016) in R (<https://www.R-project.org/>) (R Core Team 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, that is, taking their degree of clustering into account; Z-score analysis reveals the degree of local enrichment of overlaps (Supplemental Methods).

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 to assess the mean distance between ATAC-seq peaks and deletion breakpoints, for common and singleton variants separately.

### Brain enhancer data

Active brain enhancers came from de la Torre-Ubieta et al. (2018). Specifically, we used the 37,050 brain enhancers that 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, namely, overlapping the spermatogonial ATAC-seq peaks. To correct for the variable size of the brain active enhancers, we took the midpoints of each enhancer  $\pm 500$  bp on either side and intersected these sites with

the ATAC-seq peaks using BEDTools' "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 data set, 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 analyze the enrichment of short indels and SNPs around TFBSs and brain enhancers, we only considered genomic regions with unique mappability in  $\geq 95\%$  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, modeling 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 Consortium 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 one), and the overlap with "LTRs," "SINEs," "LINEs" and "simple repeats" (downloaded from the UCSC Table Browser at <https://genome.ucsc.edu/>).

In a smaller model, we subset the data set to only include 5-kb bins that also overlap active brain enhancers (de la Torre-Ubieta et al. 2018), and 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

To find sequence motifs within the 5- to 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  $\geq 95\%$  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 motif family (using BEDTools’ “bedtools sample”) (Quinlan and Hall 2010); circular permutations were performed to compare the observed to expected overlap of the control motif sites ( $\pm 10$  bp) with the gnomAD singleton insertions of 5–20 bp.

The FIMO predicted control sites were also used to assess the degree of “clustering” of motifs at spermatogonia active sites. For this purpose, we intersected the FIMO motifs with (1) spermatogonial ATAC-seq sites and (2) ENCODE Master regulatory sites downloaded from <https://genome.ucsc.edu/> (DNase I hypersensitivity derived from assays in 95 cell types). For each of the 167 motif families, we calculated the median distance (in base pairs) from a motif located within the active regulatory region to the nearest FIMO motif of the same type. Accordingly, the ratio of the median distance between motif sites (ENCODE/spermatogonia) was larger than one if motifs at spermatogonial sites were, on average, closer to each other than motifs near ENCODE sites, and we used this ratio as a measure of motif clustering. When correlating the IFE with the degree of motif clustering (Fig. 6B), we thus largely correct for base compositional biases near active sites (which impact mutation rates) (Supplemental Fig. S9), as well as the effects of historical selection on the clustering of motifs near genes; that is, shorter inter-motif distances in spermatogonia indicate that these sites have specifically high levels of motif density in spermatogonia, beyond the levels expected for binding sites in general.

## Data access

All raw sequencing data generated in this study have been submitted to the European Genome-phenome Archive (EGA; <https://ega-archive.org/>) under accession number EGAS00001005366. The ATAC-seq peak files generated in this study are available as Supplemental Data (Supplemental Datasets 1–3) and at Edinburgh DataShare (<https://doi.org/10.7488/ds/3053>).

## Competing interest statement

The authors declare no competing interests.

## Acknowledgments

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 centers; 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 thank 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. We thank Roland Donat and C.J. Shukla for tissue sam-

ple collection, Elisabeth Freyer for assistance with the FAC sorting, and Wendy A. Bickmore for useful comments to the manuscript. Approved researchers can obtain the SSC population data set described in this study by applying at <https://base.sfari.org>. This work was supported by Medical Research Council (MRC) Human Genetics Unit core funding program grants MC\_UU\_00007/11, MC\_UU\_00007/2, and MC\_UU\_00007/16.

**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.

## References

- The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073. doi:10.1038/nature09534
- The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74. doi:10.1038/nature15393
- Afek A, Shi H, Rangadurai A, Sahay H, Senitzki A, Khani 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. doi:10.1038/s41586-020-2843-2
- 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**: e28383. doi:10.7554/eLife.28383
- 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**: eaat6576. doi:10.1126/science.aat6576
- 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. doi:10.1038/nature12787
- 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–W208. doi:10.1093/nar/gkp335
- 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. doi:10.1126/science.1183439
- 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. doi:10.1038/ng.658
- 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. doi:10.1038/ncomms6969
- 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. doi:10.1371/journal.pgen.1006315
- 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. doi:10.1038/nmeth.2688
- 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**: 3573. doi:10.1038/s41467-018-05936-5
- Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg 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. doi:10.1093/bioinformatics/btv710
- 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. doi:10.1186/s13059-017-1158-6
- Collins RL, Brand H, Karczewski KJ, Zhao X, Alfoldi 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. doi:10.1038/s41586-020-2287-8
- 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. doi:10.1038/nature08516
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158. doi:10.1093/bioinformatics/btr330
- Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* **10**: giab008. doi:10.1093/gigascience/giab008
- 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. doi:10.1093/nar/gkx1081
- Deciphering Developmental Disorders Study. 2015. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**: 223–228. doi:10.1038/nature14135
- 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.e18. doi:10.1016/j.cell.2017.12.014
- The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74. doi:10.1038/nature11247
- 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. doi:10.1093/bioinformatics/btv562
- The Genome of the Netherlands Consortium. 2014. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46**: 818–825. doi:10.1038/ng.3021
- 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. doi:10.1038/s41467-020-17305-2
- 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. doi:10.1371/journal.pgen.1002334
- Gonzalez-Perez A, Sabarinathan R, Lopez-Bigas N. 2019. Local determinants of the mutational landscape of the human genome. *Cell* **177**: 101–114. doi:10.1016/j.cell.2019.02.051
- 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. doi:10.1101/gr.104471.109
- Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**: 1017–1018. doi:10.1093/bioinformatics/btr064
- Grey C, Barthes P, Chauveau-Le Fricq E, 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**: e1001176. doi:10.1371/journal.pbio.1001176
- Gu W, Zhang F, Lupski JR. 2008. Mechanisms for human genomic rearrangements. *Pathogenetics* **1**: 4. doi:10.1186/1755-8417-1-4
- 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.e6. doi:10.1016/j.stem.2017.09.003
- 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. doi:10.1038/s41422-018-0099-2
- 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**: eaau1043. doi:10.1126/science.aau1043
- 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**: 170–175. doi:10.1038/nature10336
- Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* **12**: 756–766. doi:10.1038/nrg3098
- Jiang P, Singh M. 2014. CCAT: combinatorial code analysis tool for transcriptional regulation. *Nucleic Acids Res* **42**: 2833–2847. doi:10.1093/nar/gkt1302
- 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. doi:10.1038/nature24018
- 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. doi:10.1186/s13059-018-1483-4
- Kaiser VB, Taylor MS, Semple CA. 2016. Mutational biases drive elevated rates of substitution at regulatory sites across cancer types. *PLoS Genet* **12**: e1006207. doi:10.1371/journal.pgen.1006207
- 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. doi:10.1038/s41586-020-2308-7
- Karimzadeh M, Ernst C, Kundaje A, Hoffman MM. 2018. Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Res* **46**: e120. doi:10.1093/nar/gky677
- 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. doi:10.1186/s13059-019-1894-x
- Kim S, Peterson SE, Jasin M, Keeney S. 2016. Mechanisms of germ line genome instability. *Semin Cell Dev Biol* **54**: 177–187. doi:10.1016/j.semcdb.2016.02.019
- Kondrashov AS, Rogozin IB. 2004. Context of deletions and insertions in human coding sequences. *Hum Mutat* **23**: 177–185. doi:10.1002/humu.10312
- Koren A, Polak P, Nemes 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. doi:10.1016/j.ajhg.2012.10.018
- Kribelbauer JF, Rastogi C, Bussemaker HJ, Mann RS. 2019. Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes. *Annu Rev Cell Dev Biol* **35**: 357–379. doi:10.1146/annurev-cellbio-100617-062719
- Kvikstad EM, Tyekucheva S, Chiaromonte F, Makova KD. 2007. A macaque's-eye view of human insertions and deletions: differences in mechanisms. *PLoS Comput Biol* **3**: 1772–1782. doi:10.1371/journal.pcbi.0030176
- Leppa VM, Kravitz SN, Martin CL, Andrieux J, Le Caignec C, Martin-Coignard D, DyBuncio C, Sanders SJ, Lowe JK, Cantor RM, et al. 2016. Rare inherited and *de novo* CNVs reveal complex contributions to ASD risk in multiplex families. *Am J Hum Genet* **99**: 540–554. doi:10.1016/j.ajhg.2016.06.036
- Levinson G, Gutman GA. 1987. High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. *Nucleic Acids Res* **15**: 5323–5338. doi:10.1093/nar/15.13.5323
- Li C, Luscombe NM. 2020. Nucleosome positioning stability is a modulator of germline mutation rate variation across the human genome. *Nat Commun* **11**: 1363. doi:10.1038/s41467-020-15185-0
- Li ZJ, Schulz MH, Look T, Begemann M, Zenke M, Costa IG. 2019. Identification of transcription factor binding sites using ATAC-seq. *Genome Biol* **20**: 45. doi:10.1186/s13059-019-1642-2
- Lieber MR, Ma Y, Pannicke U, Schwarz K. 2003. Mechanism and regulation of human non-homologous DNA end-joining. *Nat Rev Mol Cell Biol* **4**: 712–720. doi:10.1038/nrm1202
- MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. 2014. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* **42**: D986–D992. doi:10.1093/nar/gkt958
- Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. 2019. Structural variant calling: the long and the short of it. *Genome Biol* **20**: 246. doi:10.1186/s13059-019-1828-7
- Makova KD, Yang S, Chiaromonte F. 2004. Insertions and deletions are male biased too: a whole-genome analysis in rodents. *Genome Res* **14**: 567–573. doi:10.1101/gr.1971104
- Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* **17**: 10–12. doi:10.14806/ej.17.1.200
- McRae JF, Clayton S, Fitzgerald TW, Kaplanis J, Prigmore E, Rajan D, Sifrim A, Aitken S, Akawi N, Alvi M, et al. 2017. Prevalence and architecture of *de novo* mutations in developmental disorders. *Nature* **542**: 433–438. doi:10.1038/nature21062
- McVean G. 2010. What drives recombination hotspots to repeat DNA in humans? *Philos Trans R Soc Lond B Biol Sci* **365**: 1213–1218. doi:10.1098/rstb.2009.0299

- Messer PW. 2009. Measuring the rates of spontaneous mutation from deep and large-scale polymorphism data. *Genetics* **182**: 1219–1232. doi:10.1534/genetics.109.105692
- Messer PW, Arndt PF. 2007. The majority of recent short DNA insertions in the human genome are tandem duplications. *Mol Biol Evol* **24**: 1190–1197. doi:10.1093/molbev/msm035
- Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, et al. 2012. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**: 1431–1442. doi:10.1016/j.cell.2012.11.019
- Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS, et al. 2013. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res* **23**: 749–761. doi:10.1101/gr.148718.112
- Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**: 321–324. doi:10.1126/science.1117196
- Myers S, Freeman C, Auton A, Donnelly P, McVean G. 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet* **40**: 1124–1129. doi:10.1038/ng.213
- Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G, Donnelly P. 2010. Drive against hotspot motifs in primates implicates the *PRDM9* gene in meiotic recombination. *Science* **327**: 876–879. doi:10.1126/science.1182363
- Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, et al. 2012. BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**: 1919–1920. doi:10.1093/bioinformatics/bts277
- Palmer N, Talib SZA, Ratnacaram CK, Low D, Bisteau X, Lee JHS, Pfeifferberger E, Wollmann H, Tan JHL, Wee S, et al. 2019. CDK2 regulates the NRF1/Ehmt1 axis during meiotic prophase I. *J Cell Biol* **218**: 2896–2918. doi:10.1083/jcb.201903125
- Powers NR, Parvanov ED, Baker CL, Walker M, Petkov PM, Paigen K. 2016. The meiotic recombination activator PRDM9 trimethylates both H3K36 and H3K4 at recombination hotspots in vivo. *PLoS Genet* **12**: e1006146. doi:10.1371/journal.pgen.1006146
- Pratto F, Brick K, Khil P, Smagulova F, Petukhova GV, Camerini-Otero RD. 2014. DNA recombination. recombination initiation maps of individual human genomes. *Science* **346**: 1256442. doi:10.1126/science.1256442
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842. doi:10.1093/bioinformatics/btq033
- Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dünder F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**: W160–W165. doi:10.1093/nar/gkw257
- Ramu A, Noordam MJ, Schwartz RS, Wuster A, Hurles ME, Cartwright RA, Conrad DF. 2013. DeNovoGear: de novo indel and point mutation discovery and phasing. *Nat Methods* **10**: 985–987. doi:10.1038/nmeth.2611
- R Core Team. 2016. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. <https://www.R-project.org/>.
- Reijns MAM, Kemp H, Ding J, de Procé SM, Jackson AP, Taylor MS. 2015. Lagging-strand replication shapes the mutational landscape of the genome. *Nature* **518**: 502–506. doi:10.1038/nature14183
- Rodgers K, McVey M. 2016. Error-prone repair of DNA double-strand breaks. *J Cell Physiol* **231**: 15–24. doi:10.1002/jcp.25053
- Sabarinathan R, Mularoni L, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. 2016. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**: 264–267. doi:10.1038/nature17661
- Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32**: D91–D94. doi:10.1093/nar/gkh012
- Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V, Jaakkola T, Gifford DK. 2014. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol* **32**: 171–178. doi:10.1038/nbt.2798
- Short PJ, McRae JF, Gallone G, Sifrim A, Won H, Geschwind DH, Wright CF, Firth HV, FitzPatrick DR, Barrett JC, et al. 2018. De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555**: 611–616. doi:10.1038/nature25983
- 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.e4. doi:10.1016/j.celrep.2019.01.045
- 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. doi:10.1038/ng.363
- 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. doi:10.1016/j.ajhg.2015.11.023
- 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.e12. doi:10.1016/j.cell.2017.08.047
- van Gent DC, Hoeijmakers JH, Kanaar R. 2001. Chromosomal stability and the DNA double-stranded break connection. *Nat Rev Genet* **2**: 196–206. doi:10.1038/35056049
- 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. doi:10.1096/fj.201700093R
- Wellcome Trust Case Control Consortium, 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 3000 shared controls. *Nature* **464**: 713–720. doi:10.1038/nature08979
- 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. doi:10.1038/s41588-018-0107-y
- Wolfe KH, Sharp PM, Li WH. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* **337**: 283–285. doi:10.1038/337283a0
- 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. doi:10.1186/s13059-020-1929-3
- 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. doi:10.1038/npjgenmed.2016.27
- Yuen RKC, Merico D, Bookman M, Howe JL, Thiruvahindrapuram B, Patel RV, Whitney J, Deflaux N, Bingham J, Wang Z, et al. 2017. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat Neurosci* **20**: 602–611. doi:10.1038/nn.4524
- Zhang Y, Liu T, Meyer CA, Eeckhoutte J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based Analysis of CHIP-Seq (MACS). *Genome Biol* **9**: R137. doi:10.1186/gb-2008-9-9-r137

Received February 22, 2021; accepted in revised form August 19, 2021.