

Genome-wide annotation of human multi-nucleotide variants reveals widespread functional differences from single nucleotide variants

Corresponding Author: Professor Jing Gong

Parts of this Peer Review File have been redacted as indicated to remove third-party material.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Jin et al provided an extended catalog of multi nucleotide variants (MNVs) together with rich biological analysis and browser visualizations.

The work is highly valuable and is going to be a great resource. Some additional analysis as mentioned below might be needed to enhance its value:

Major comments:

- The fact that MNVs are enriched for particular regions/contexts could imply two different things:

1. Simply due to increase of SNV rate

2. Increase of MNVs relative to SNV rate

and 2. are more likely to be pol-zeta error / repeat slippage etc.

Could the author try to differentiate these two biologies? This is well done in the African population-specific aspect of MNV analysis (As the author mention in lines 113-114).

For other analysis such as functional annotation enrichment (lines 160-163), I recommend

* defining MNV density normalized by SNV density to differentiate two scenarios above, and/or

* have a comparison of the rates for different MNV categories (e.g. SNV combination vs pol-zeta/slippage) compared to baseline

* and have a brief discussion on their implications

to better demonstrate that the MNV enrichment we are observing is a natural consequence of higher SNV mutation rate, or something beyond and specific to MNVs.

- Could the author provide additional clarification where the difference in the MNVs identified per method is coming from?

Since the definition of MNV itself is not ambiguous, ideally all tools should identify the same set of "true" MNVs.

Reading the manuscript "Generation of the gold standard dataset" and other sections, seems like the differences are possibly coming from read depth cutoff, ability to merge sub-joint MNVs to higher joint MNVs, and/or some additional filtering criteria? Same for the precision/recall, could the authors mention why some methods are very low in precision/recall and why the proposed new method still does not achieve 100% precision/recall? This is clearly mentioned for BCFtools and MACARON (lines 90-92) but not clear enough for other tools.

- I appreciate that the authors mention some of their findings are consistent with prior literature such as Wang et al (which is reassuring), but such comparison with existing literature seems to be missing for some part?

For example,

* the fact that MNVs are depleted around TSS, probably due to differences in GC content, is mentioned in Wang et al.

* The "Mutational patterns of different MNV types" overall seems to be consistent with Wang et al. Does the percentage

roughly match?

* The MNV consequence in coding regions as in line 164-177 were also well explored in Wang et al. Are these findings described here in a good alignment? Are the MNVs the authors reporting here novel, not covered in gnomAD (Wang et al) for particular reasons?

Did the author find additional features for these sections beyond existing literature?

- I recognize one of the key uniqueness of this work compared to existing ones is the functional annotation of non-coding regions, defining TFBS/miRBS creating/disrupting MNVs.

These are really interesting (and the experimental result is beautiful!) but could be further analyzed by somewhat orthogonal aspects. For example,

* For TFBS creating/disrupting MNVs, do we see these to be disrupted for particular class of TFs? Does the MNV rate reflect constraint of TFs and/or the target genes? For example, in Wang et al, gained nonsense MNVs were disrupted for constrained genes, which provided an orthogonal evidence that gained nonsense MNVs are more functionally impactful compared to other classes. Could the author provide similar discussion suggesting that TFBS MNVs are indeed functionally impactful?

* For non-coding regions especially, GWAS association itself often times does not imply causality, as there are many variants in linkage (LD), unless statistically fine-mapped with high PIP.

Do the author deem the miRNA MNVs associated in GWAS to be causal? Fine-mapping the associated loci or at least checking the locus zoom plot to see whether the MNV-constituting SNV is the lead variant would be helpful.

- Fig 5a is very interesting comparing the delta MFE, but the real key question is to see whether the delta MFE of an MNV is more/less than the sum of delta MFE of the constituent SNV1 and SNV2.

This is analogous to, for example, gained nonsense MNVs where the combined effect is clearly beyond the summation of individual SNVs.

Also, the experimental result the authors provide is a beautiful example where the effect is beyond additive.

Could the author make such comparisons for each miRBS MNVs and discuss MNVs effect beyond/below additive?

Minor comments:

- Line 85 "high" and "low" MNVs should be clearly defined.

- Line 106 and other places: GnomAD should be gnomAD.

- Fig 2d and several other figures: Venn diagram with more than 3 entries is very hard to visualize. Could the authors use upsetr plot instead?

- Fig. 2e: x and y axis should be PC1 and PC2 rather than "PCA" 1 and 2? (A stands for "Analysis" and should not be needed here for the legend).

- Line 120-127: The biological connection here is a bit superficial. Is the number "12 SNVs" more than expected by chance? I would not mind deleting this entire discussion, or providing some additional details.

- Line 161-163: This MNV density enrichment could probably be explained simply by SNV mutation rate differences?

- Fig. 5h experiment: Why is the NC group somewhere in between WT and MNV? Is this expected?

- Line 246-247: Is this more than expected by chance? In other words, are these MNVs enriched in oncogenes, compared to random MNVs?

- line 250: Should clarify that $p < 0.05$ is "nominally" significant, or apply multiple test correction.

(Remarks on code availability)

The webpage is a nice browser but I don't see the source code (maybe I am missing it)

Reviewer #2

(Remarks to the Author)

The manuscript "Genome-wide identification and annotation of multi-nucleotide variants in human reveal widespread functional differences compared to single nucleotide variants" presents a large-scale analysis of multi-nucleotide variants (MNVs) across over 700,000 human samples. The authors develop MNVAnno, a new tool for identifying and annotating MNVs, and compile the largest human MNV catalog to date. They demonstrate that MNVs can have distinct functional consequences compared to single nucleotide variants (SNVs), affecting both coding and non-coding regions, and identify potential disease associations, including with several cancers. The work provides valuable resources and an elegant analysis, but certain methodological and interpretative aspects require clarification and strengthening.

Comments

Definition and complexity of MNVs

In the introduction, the manuscript refers to "complex MNVs" but does not define what qualifies as complex. Additionally, a more detailed explanation of the challenges in MNV annotation is needed. This would set the context for why existing tools struggle and what precisely MNVAnno improves.

Sources of false positives and false negatives

The introduction or Methods should discuss in more depth the sources of error in MNV annotation. What leads to false positives or false negatives? Is it related to phasing quality only, or other factors are important?

Normalization for sample size and mutation rate (Figure 4a)

The authors compare MNV densities across datasets and cohorts, but do not correct for differences in sample size or mutation rate. Models such as the Roulette or Chen et al. (Nature 2024) mutation rate model could be used for normalization. Without this, density differences between cohorts reflect mainly differences in sample size, which is not an interesting feature. More importantly, the density difference between genomic categories can be caused by mutation rate differences.

Cardiomyopathy association (lines 120–126)

The enrichment of MNVs in cardiomyopathy pathways in African populations is interesting but should be compared to the same analysis using only SNVs. It is unclear whether this signal reflects truly MNV-specific effects or the aggregation of constituent SNVs already known from GWAS.

Additivity of functional effects (Figure 5a)

Figure 5a shows that MNVs produce larger changes in RNA secondary structure ($|\Delta\text{MFE}|$) than SNVs, but the crucial question is whether these effects are non-additive. The authors should compare the MNV effect to the sum of the effects of its constituent SNVs to demonstrate epistatic or synergistic interactions—this is central to the biological rationale for analyzing MNVs as combined events.

GWAS significance threshold

The use of $p < 5 \times 10^{-6}$ as a genome-wide threshold (line 246) can be too permissive. The standard threshold for GWAS is 5×10^{-8} , or alternatively, an appropriate Bonferroni or FDR correction should be applied based on the number of MNVs tested. Without this correction, the reported 267 associations may include false positives.

This manuscript presents valuable resources and advances MNV research by extending annotation to complex and non-coding variants. However, several conceptual and methodological issues should be clarified before publication.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I truly appreciate the authors' effort addressing my questions. Most of my original questions are well addressed. I hope the following three remaining points could be further addressed before publication. Thank you so much for the authors' effort in advance.

In brief, I hope the authors could:

- 1) Re-visit the interpretation of (MNV/SNV) ratio
- 2) Check whether difference with gnomAD MNVs could be explained by inclusion of low complexity region, and
- 3) Provide additional interpretation of deviation of delta MFE from additive

Below are some more details:

1) Re-visit the interpretation of (MNV/SNV) ratio:

I think the unit of MNV/SNV ratio is a bit tricky - for one-step event, normalization by SNV rate is fine, but strictly speaking, for two-step event, MNV is a combination of SNVs, so the normalization should be dividing by the square of SNV rate (i.e. SNV rate ^{2} , sorry that I did not clarify this in my original comment. MNV rate should also reflect how uneven the SNVs are distributed, but that is the kind of complexity beyond our analysis.). Probably that reflects the fact that the sharp curve of Type "SNV event" versus relatively flat curve of other one-step type events in Figure R1.

Probably due to this fact, the MNV/SNV ratio might not fully reflect conservation. For example,

- exons and splice regions are in general more conserved than TFBS, but the MNV/SNV ratio does not reflect that.
- 5'UTR should be relatively conserved but as the authors mentioned, we don't see depletion of MNV/SNV ratio because of differences of CpG density which makes the SNV rate different.

The 5'UTR observation makes sense since, let the global SNV rate be p and SNV rate in a particular functional category be

$c \cdot p$. For a two step MNV event, ignoring the positional dependencies for simplicity, MNV rate would scale with $(c \cdot p)^2$, so (MNV/SNV) would be $(c \cdot p)^2 / (c \cdot p) = c \cdot p$. This means that if the mutation rate itself is high in the category (i.e. $c > 1$), the (MNV/SNV) is also going to be high, for two-step MNV events. This could probably explain the fact that 5'UTR, rich in CpG methylation, has high (MNV/SNV) . This also seems to match Fig. 2c, where visually $y = x^2$ could be a better fit.

Therefore, the logic that “low MNV/SNV => stronger evolutionary constraint specifically on MNV” may not hold true. I think we need to be careful about the description in the newly added sentences: “This reflects an evolutionary constraint against MNVs in these functionally critical regions, indicating that MNVs are more deleterious than individual SNVs from the evolutionary perspective” - this could be true evolutionary constraint OR just the limitation of the way we normalize the MNV rate.

Given these facts, here are two additional recommendations:

1. Look at the $\text{MNV}/(\text{SNV}^2)$ ratio in addition to the current MNV/SNV ratio, and check which parts of the discussion still hold true.
2. Lower the tone of the sentence above. Instead of mentioning “this reflects an evolutionary constraint”, the authors could write something like “this *could* reflect.., *although* it could simply reflect the complex relationship between SNV rate vs MNV rate [ref. 22]”.

If the authors think 1. is beyond the scope of the paper, just 2. alone is fine.

2) Check whether difference with gnomAD MNVs could be explained by inclusion of low complexity region

I doubt that “This discrepancy arises primarily because our study incorporates the 1000G project (with samples from five populations), whereas gnomAD is dominated by European populations.”, as the gnomAD v2 dataset also includes a good number of non-european samples (e.g. it contains >9,000 East Asian samples, far more than that in 1000G <https://www.nature.com/articles/s41586-020-2308-7/figures/1>).

Could this be rather due to differences in the variant filtering? In particular, considering that in gnomAD, MNVs in low complexity regions were conservatively filtered out (as stated in the manuscript <https://www.nature.com/articles/s41467-019-12438-5>: “These findings come with the caveat that variants in repetitive regions will have higher error rates due to slippage and misalignment errors, but we have reduced this risk by applying random forest filtering for individual sites, as well as removing all the variants in low-complexity regions from our analysis”), I wonder a large majority of those MNVs uniquely discovered in this study are actually in low complexity regions and are filtered out in gnomAD (or just generally relatively low-quality variants that are filtered out in gnomAD).

Could the authors check whether this is the case? If it is, that is still fine as long as that is clearly mentioned. Other than that, I fully agree with the added value of 3-5 joint MNVs.

3) Provide additional interpretation of deviation of delta MFE from additive

Figure R10 is beautiful. Is there a way to quantify whether these deviations are more than expected by chance? If that is hard (maybe it is, since we cannot get the null distribution), I think mentioning the direction of the non-additive effect in Figure R10 is meaningful here, to show that it is not random and to interpret it biologically. Specifically, I see the mitigation effect (i.e. when sum of SNV effects are large, MNV effects tend to be smaller than the sum, and vice versa), so in many cases, we expect $(\text{effect of one SNV}) < (\text{effect of MNV}) < (\text{sum of the effects of constituent SNVs})$, if I am interpreting the figure correctly. Mentioning this observation together with an interpretation would be valuable.

(Remarks on code availability)
Code is fine.

Reviewer #2

(Remarks to the Author)

The authors have responded to my comments and implemented meaningful improvements to the article. The results about the non-additivity are particularly interesting.

One minor issue is the inconsistency in the number of significant MNVs reported (232 in the response vs. 233 in the revised manuscript), which should be clarified and corrected.

Once this is clarified, I support acceptance for publication.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I truly appreciate the authors' additional efforts. My questions are well addressed. I have one final comment regarding my previous question "2) Check whether difference with gnomAD MNVs could be explained by inclusion of low complexity region", as detailed below.

If the authors agree on the interpretation, adding/editing a few sentences in the main text before final publication would be appreciated.

I do not have other additional concerns/comments. Thank you so much again for working on all the revisions and congratulations on the completion of this amazing work!

Best,
Qingbo Wang

Regarding "2) Check whether difference with gnomAD MNVs could be explained by inclusion of low complexity region":

Thanks a lot for performing additional analysis to identify that the inclusion/exclusion of low complexity regions (LCRs) explains a significant amount of the difference in the number of MNVs identified in gnomAD vs this study. While I agree that including LCRs to construct a more comprehensive list of MNVs is a rational choice, I think the fact should be more explicitly stated in the main text and methods, not only in the supplementary text.

I appreciate it if the authors could add a few sentences clarifying the facts that

- Approximately 70% of the "novel" joint MNVs are in LCRs
- LCRs are known to be enriched for sequencing errors, so the "novel" joint MNVs could contain some fraction of false-positives
- The users should take caution when interpreting these joint MNVs in LCRs
- (but the authors included them nevertheless, to make the list more comprehensive)

Although I will let the authors decide where and how to include these statements, it seems that the places such statement could be included are around:

- Lines 134-135 "...2,222,090 135 MNVs were newly discovered compared to those in gnomAD..."
- Lines 365-366 "...found 2,222,090 novel ones compared with previous Wang's study utilizing gnomAD dataset..."
- Lines 410 "Identification and annotation of MNVs" section of the method
- Figure 2a legend

I do agree that other factors such as sample composition, inclusion of 3- to 5- joint MNVs are also contributing factors. Just want to make sure that the readers understand that different filtering criteria is one of the main sources of the difference in the number of MNVs discovered, and won't get over-excited about the fact that we are discovering so many additional MNVs. Thanks a lot.

(Remarks on code availability)

Code is fine.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Point-by-point response to the reviewers' comments

Reviewer #1:

Jin *et al.* provided an extended catalog of multi nucleotide variants (MNVs) together with rich biological analysis and browser visualizations.

The work is highly valuable and is going to be a great resource. Some additional analysis as mentioned below might be needed to enhance its value.

Response: Thank you very much for the time and thoughtful comments you devoted to reviewing our manuscript. We sincerely appreciate your positive assessment of our work and recognition of its value to the research community. We have carefully addressed all raised concerns with additional analyses, and revised the manuscript accordingly. Please see the point-to-point response below.

Major comments:

1. The fact that MNVs are enriched for particular regions/contexts could imply two different things: 1. Simply due to increase of SNV rate. 2. Increase of MNVs relative to SNV rate. and 2. are more likely to be pol-zeta error / repeat slippage etc.

Could the author try to differentiate these two biologies? This is well done in the African population-specific aspect of MNV analysis (As the author mention in lines 113-114).

For other analysis such as functional annotation enrichment (lines 160-163), I recommend defining MNV density normalized by SNV density to differentiate two scenarios above, and/or have a comparison of the rates for different MNV categories (e.g. SNV combination vs pol-zeta/slippage) compared to baseline and have a brief discussion on their implications, to better demonstrate that the MNV enrichment we are observing is a natural consequence of higher SNV mutation rate, or something beyond and specific to MNVs.

Response: Thank you very much for this insightful and constructive comment, which helps us better demonstrate whether enrichment is a consequence of higher SNV rate or MNV-specific mechanisms.

To directly address this question, we performed two complementary analyses.

(1) MNV/SNV ratio analysis across functional genomic regions

Following your suggestion, we employed the MNV/SNV ratio to correct for the influence of SNV rates. Using whole-genome sequencing (WGS) data from 1000G and GTEx datasets, we calculated the MNV/SNV Ratio (MNV density normalized by SNV density) for all functional genomic regions, with the genome-wide average ratio (the number of all MNVs / the number of all SNVs) set as the baseline. As shown in Figure R1a-b, we observed that in most regions (e.g., intron, lncRNA, miRNA, 3'UTR), the MNV/SNV ratios are around the baseline, indicating that MNV occurrence in these regions largely parallels the SNV mutation rate. Highly conserved regions (e.g., exon, splice, conserved element, miRBS, and TFBS) exhibited significantly lower MNV/SNV ratios. This reflects an evolutionary constraint against MNVs in these functionally critical regions, indicating that MNVs are more deleterious than individual SNVs from the evolutionary aspect. Interestingly, after normalization, we observed that the 5'UTR region showed a relative enrichment of MNVs. Prior studies have indicated that the 5'UTR region is rich in CpG dinucleotides^{1,2}, which are more likely to induce the generation of MNVs³.

(2) Distribution of three MNV types across genomic regions

We further stratified MNVs into three mechanistic categories (SNV event, pol-zeta-mediated one-step, and repeat slippage) based on the framework established by Wang *et al.*³ and compared their distribution across genomic regions. As shown in Figure R1c-d, the MNV/SNV patterns of three mechanistic categories were similar to the overall profile in Figure 1a-b. In summary, most regions parallel with the baseline SNV mutation rates, while highly conserved regions exhibit reduced ratios, indicative

of evolutionary constraint. The 5'UTR again shows relative MNV enrichment, especially for SNV events and pol-zeta-mediated one-step. This subtle divergence may be attributed to the high abundance of CpG dinucleotides in the 5'UTR ^{1,2}: MNVs classified as SNV events tend to be enriched at CpG transition sites ³⁻⁶, whereas GC→AA changes and their reverse complements represent a major substitution type in pol-zeta-mediated one-step events ⁷⁻¹⁰. Thus, compared with MNVs resulting from repeat slippage, the enrichment of the other two mechanistic categories in the 5'UTR can be partially explained by the sequence features of this region and their correspondence with specific mutational mechanisms.

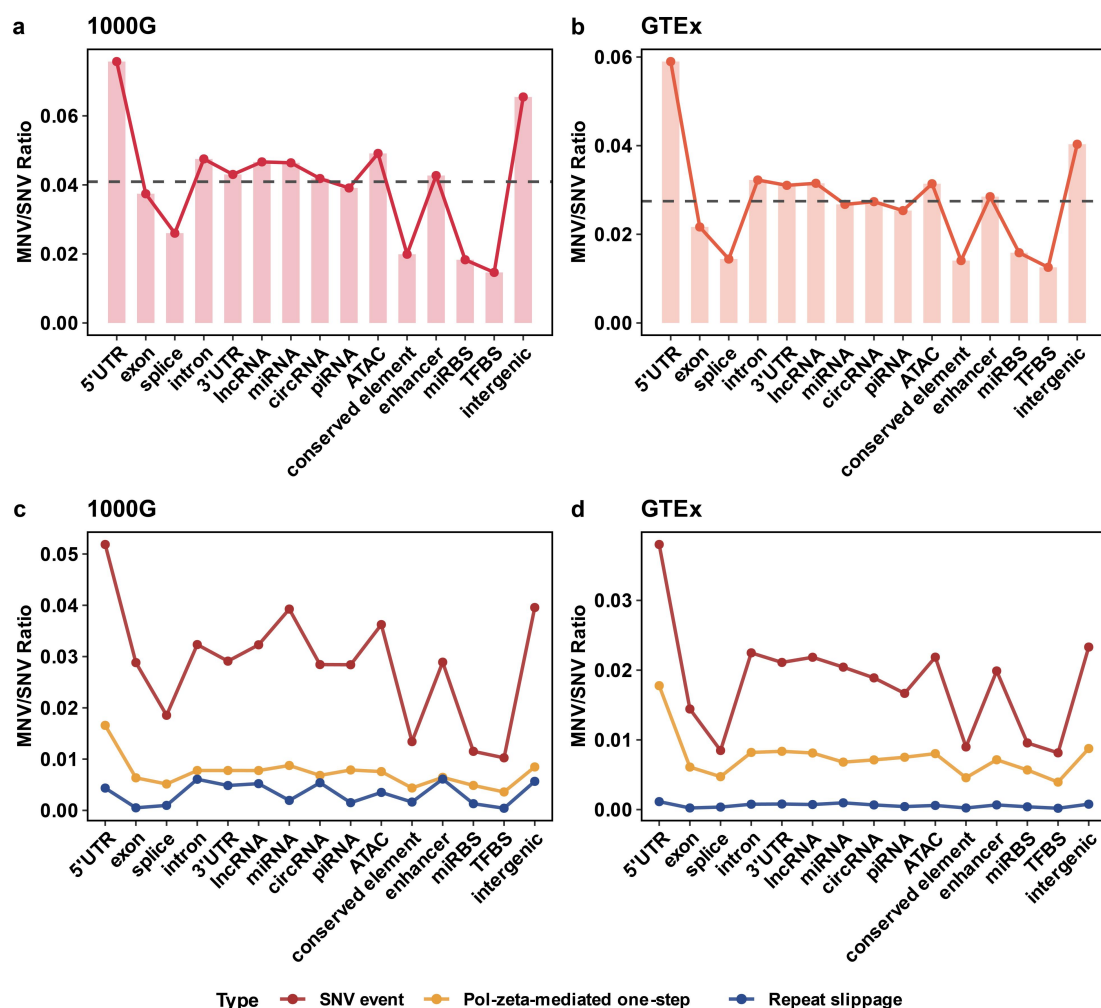


Figure R1. The distribution of the MNV/SNV ratios in different regions of 1000 Genomes Project and GTEx. (a,b) MNV/SNV ratios across annotated genomic regions in the 1000 Genomes Project (1000G; a) and GTEx (b). The dashed horizontal

line represents the genome-wide background (baseline) MNV/SNV ratio. (c,d) MNV/SNV ratios of three mechanistic categories (SNV event, pol-zeta-mediated one-step, and repeat slippage) in 1000G (c) and GTEx (d).

We have integrated these results and Response figure into the revised manuscript. In the revised manuscript, we have added the following sentences in lines 195-208.

“After normalizing the genomic length, a similar density distribution was observed across whole-genome sequencing datasets (Fig. 4a). Moreover, to correct for regional differences in SNV mutation rate, we calculated the MNV/SNV ratio (MNV density normalized by SNV density) for all functional genomic regions in WGS datasets. We observed that in most regions (e.g., intron, lncRNA, miRNA, 3'UTR), the MNV/SNV ratios were around the baseline, indicating that MNV occurrence in these regions largely paralleled the SNV mutation rate (Supplementary Fig. 9). Highly conserved regions (exon, splice region, conserved element, miRBS, and TFBS) exhibited significantly lower MNV/SNV ratios. This reflects an evolutionary constraint against MNVs in these functionally critical regions, indicating that MNVs are more deleterious than individual SNVs from the evolutionary perspective. Notably, after normalization, we observed that the 5'UTR region exhibited a relative enrichment of MNVs (Supplementary Fig. 9). Previous studies have indicated that the 5'UTR region is rich in CpG dinucleotides^{29,30}, which are more likely to promote the generation of MNVs⁹.”

2. Could the author provide additional clarification where the difference in the MNVs identified per method is coming from? Since the definition of MNV itself is not ambiguous, ideally all tools should identify the same set of "true" MNVs.

Reading the manuscript "Generation of the gold standard dataset" and other sections, seems like the differences are possibly coming from read depth cutoff, ability to merge sub-joint MNVs to higher joint MNVs, and/or some additional filtering criteria?

Same for the precision/recall, could the authors mention why some methods are very low in precision/recall and why the proposed new method still does not achieve 100% precision/recall? This is clearly mentioned for BCFtools and MACARON (lines 90-92) but not clear enough for other tools.

Response: Thank you very much for this comment. To address this, we conducted a further careful analysis of the comparison results, and found that the differences in MNV identification results among various methods stem primarily from three aspects: supported MNV types and identification scope, input data requirements, and sub-joint variant merging capability.

(1) Supported MNV types and identification scope

MNVs are defined as the coexistence of two or more nucleotide variants on the same haplotype within an individual. As shown in Figure R2a, their combinations are diverse and complex, including 2-joint and multi-joint MNVs (≥ 3 -variants) spanning coding and non-coding regions, as well as adjacent continuous and non-continuous variants.

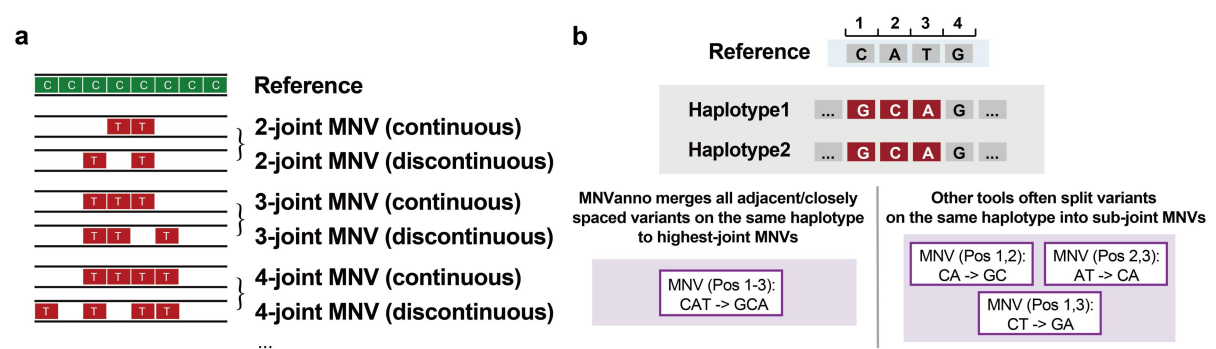


Figure R2. Types and identification strategy of MNVs. (a) Illustrative examples of the different MNV types; (b) Sub-joint MNV merging strategy

We summarized the supported MNV types and genomic scope of six tools in Table R1. Briefly, VarDict only detects continuous MNVs; Wang's tool mainly supports 2-joint MNVs and a few 3-joint MNVs within coding regions; MACARON is restricted to single codons; BCFtools focuses on exonic MNVs in intron-free General Feature

Format (GFF) regions; MAC supports multi-joint MNVs at the genome level, but it identified far fewer MNVs than other tools. In contrast, MNVAnno can handle both continuous and discontinuous ones (~80% of all MNVs), and supports 3-5 joint MNVs (~10% of our results) at the genome-wide level.

Table R1. Comparison of the application scope of six tools

Tools	Input data type	Support multiple samples	Identification scope	Supported MNV types	Annotation scope
VarDict	BAM	No	Genome	Only support continuous multi-joint MNVs	Not support
MAC	SNP (VCF) BAM	No	Genome	Support multi-joint MNVs	Function annotation in coding regions
MACARON	VCF	No	Codon	Support MNVs within a single codon	Function annotation in coding regions
BCFtools	Phased VCF	Yes	Intron-free (GFF)	Support multi-joint MNVs	Not support
Wang's tool	Phased VCF	Yes	Genome	Support 2-joint MNVs	Function annotation in coding regions and position annotation in other regions
MNVAnno	Phased VCF	Yes	Genome	Support multi-joint MNVs	Function annotation in coding regions and non-coding regions (e.g. ncRNA, 3'UTR, TFBS)

(2) Input data requirements

BAM-based tools (VarDict): VarDict relies on raw sequencing reads (BAM files) to call MNVs by detecting adjacent variants in the same read. However, it cannot verify haplotype phase (cis/trans status) of variants. This leads to false positives (e.g., incorrect read mapping and haplotype phase error) and false negatives (cannot

identify non-continuous MNVs).

Phased VCF-based tools (MNVAnno, Wang's tool, BCFtools): These tools use phased VCFs to avoid phase-related errors. However, their performance depends on phasing quality—rare errors (mean error rate < 0.005 ¹¹) still introduce minor inaccuracies.

MAC requires both a SNP list from VCF files and BAM files for its operation. It infers haplotypes from reads spanning multiple SNVs, defines connected components as “Block of Mutation (BM)”, and then identifies MNVs based on BMs. MAC was published in 2015 and identifies far fewer MNVs than other tools. However, the authors have not provided specific details regarding its identification criteria or filtering thresholds, which makes it difficult to pinpoint the exact reasons for its discrepancies compared to other methods.

(3) Sub-joint MNV merging capability

MNVAnno employs a "high-joint priority" identification strategy (Figure R2b). For example, if three adjacent or closely spaced variants are detected on the same haplotype, MNVAnno directly classifies them as a single 3-joint MNV instead of splitting them into multiple 2-joint sub-sets (e.g., variants at positions 100, 102, 103 are annotated as one 3-joint MNV, not three separate 2-joint MNVs: 100+102, 102+103 and 100+103). This merging step prevents misclassification of sub-joint variants as independent MNVs and preserves the biological authenticity. In contrast, other tools lack this merging step, leading to redundant calls of overlapping sub-joint MNVs that reduce precision and recall.

However, this strategy introduces a minor trade-off in comparative evaluations across tools, and it is the main reason MNVAnno fails to achieve 100% precision/recall in gold-standard dataset comparisons. MNVAnno's merging strategy may lead to occasional false positives and false negatives due to rare phasing errors. For instance, if a 3-joint MNV identified by MNVAnno includes one variant with phasing errors (e.g., an artifact SNV), the entire 3-joint MNV results in a false positive classification.

Meanwhile, this 3-joint MNV is split into multiple sub-joint MNVs by other tools (e.g., Wang's tool), and those sub-joints have sufficient read support, MNVAnno may be marked with false negatives for failing to call those sub-joints independently. We carefully examined the comparison results of MNVAnno. There were 391 false positives and 1,436 false negatives, and most of them were caused by high-joint MNV combination. Nevertheless, we still maintain that reporting the observed haplotype as a single MNV remains biologically more faithful than enumerating redundant sub-sets, and this principle aligns with our emphasis on analyzing MNV rather than individual constituent SNVs.

In the revised manuscript, we have updated the Table 1 and incorporated relevant content into lines 91–102 and lines 116-121.

“By definition, besides 2-joint continuous MNVs, there are also various other types of MNVs, including multi-joint MNVs (≥ 3 variants) as well as adjacent continuous and non-continuous variants spanning coding and non-coding regions (Supplementary Fig. 1). Current tools primarily focus on continuous MNVs within coding regions. For example, VarDict only identifies continuous multi-joint MNVs¹⁶, MACARON is limited to MNVs within a single codon¹⁷, BCFtools focuses on exonic MNVs in intron-free General Feature Format (GFF) regions¹⁸, and Wang’s tool is primarily suitable for 2-joint MNVs⁹ (Table 1). Thus, we developed MNVAnno (See Supplementary Materials for details), a toolbox for the identification and annotation of all kinds of MNVs (Fig. 1c). MNVAnno adopts a sub-joint MNV merging strategy, thereby avoiding the possibility of mislabeling higher joint MNVs as lower joint MNVs (e.g., a 3-joint MNV is misidentified as multiple 2-joint MNVs), and provides more comprehensive annotation in both coding and non-coding regions (Table 1).”

“The discrepancies among tools mainly stem from three aspects: (i) supported MNV types (e.g., VarDict only identifies continuous MNVs, while MNVAnno covers continuous and non-continuous multi-joint MNVs); (ii) input data requirements (which may introduce alignment errors and phasing errors); and (iii) merging strategy

for high-joint MNVs (MNVA_{anno} adopts “high-joint priority” strategy to prevent splitting 3-joint MNVs into multiple 2-joint ones, a feature not available in other tools) (See Supplementary Materials for details).”

3. I appreciate that the authors mention some of their findings are consistent with prior literature such as Wang *et al.* (which is reassuring), but such comparison with existing literature seems to be missing for some part?

For example, the fact that MNVs are depleted around TSS, probably due to differences in GC content, is mentioned in Wang *et al.*

* The "Mutational patterns of different MNV types" overall seems to be consistent with Wang *et al.* Does the percentage roughly match?

* The MNV consequence in coding regions as in lines 164-177 were also well explored in Wang *et al.* Are these findings described here in a good alignment? Are the MNVs the authors reporting here novel, not covered in gnomAD (Wang *et al.*) for particular reasons?

Did the author find additional features for these sections beyond existing literature?

Response: Thank you very much for highlighting the importance of comparing our findings with prior literature. Given that Wang *et al.* have already thoroughly elucidated the mechanistic origins of MNVs, our work primarily focused on genome-wide annotation of MNVs in both coding and non-coding regions. As such, we did not perform a detailed comparison in the original manuscript. To fill this gap and uncover additional features, we have conducted the following additional analyses.

(1) MNV pattern around TSS regions (The conclusions are roughly same)

Referring to Wang *et al.*'s study, we annotated seven regions around the transcription start site (TSS) and calculated the proportion of MNVs attributable to different mutational origins. Our results showed a high degree of consistency with those

reported by Wang *et al.* (Figure R3a vs R3b).

[Figure Redacted]

(2) Mutational patterns of different MNV types (strong consistency)

Wang *et al.* characterized three core MNV mutational mechanisms (SNV event, pol-zeta-mediated one-step, repeat slippage) and their base change patterns. We also plotted the counts of different MNV base change patterns from WGS data (Figure R4a), and the overall results were consistent with Wang *et al.*'s findings (Figure R4b). For adjacent 2-joint MNVs, CA→TG was the most abundant, whereas TA→GC was the least abundant. The remaining types showed slight differences in ranking, but in general transitions outnumbered transversions.

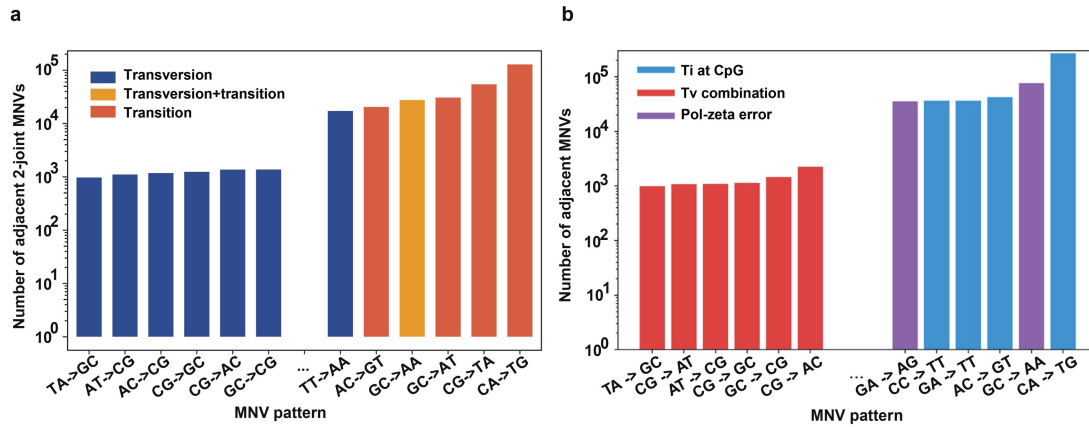


Figure R4. The number of different MNV base change patterns. The result on the left (a) is from our study, while the result on the right (b) is from Wang *et al.*'s study.

Novel insight: Wang *et al.* focused primarily on 2-joint MNVs and therefore left the patterns of higher-joint MNVs unexplored. In our study, we found that as the number of MNV joints increases (from 2 to 5), the proportion of repeat-slippage MNVs rises sharply (from 8.12% in 2-joint to 70.48% in 5-joint MNVs), accompanied by a corresponding decline in SNV-event proportions (Figure R5 and original manuscript Figure 3b). This previously unreported trend underscores the distinct patterns of high-joint MNVs.

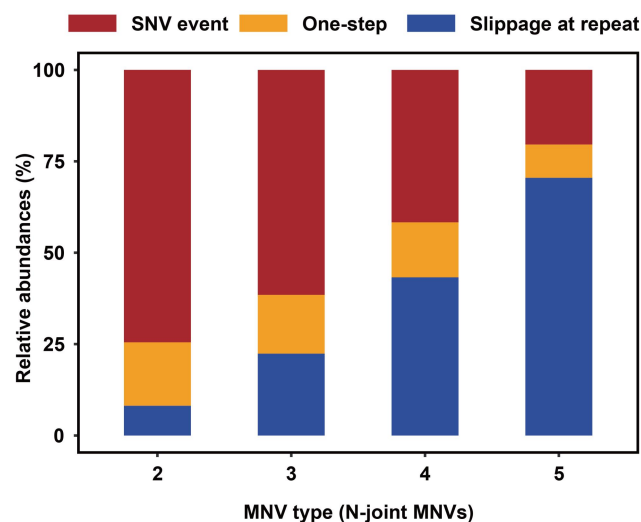


Figure R5. The proportion of three mutational categories in different MNV types.

(3) MNV consequences in coding regions (closely match)

Wang *et al.* reported that ~60% of coding MNVs lying within the same codon lead to novel amino acid changes distinct from constituent SNVs. In our results, among 37,207 2-3 joint MNVs within the same codon, 52.11% (19,388) MNVs induce novel amino acid changes—closely matching Wang *et al.*'s 60%.

(4) The novel MNVs identified in our study

Of the 3,984,258 MNVs we identified, 2,222,090 (55.78%) are not present in gnomAD (Wang *et al.*'s dataset). This discrepancy arises primarily because our study incorporates the 1000G project (with samples from five populations), whereas gnomAD is dominated by European populations. Consistent with this, 1,594,588 of these novel MNVs originate from the 1000G project, and 1,182,816 of them are from non-European populations, further underscoring the impact of population diversity on MNV discovery. Furthermore, we characterized 3-5 joint MNVs (416,409 in total) that were unreported in Wang *et al.*'s work, expanding the known repertoire of complex MNV patterns.

(5) Key novel features beyond Wang *et al.*'s work

Beyond validating Wang *et al.*'s observations, our study provides two critical novel contributions:

Complex MNV identification: We defined and systematically identified “complex MNVs” (continuous and non-continuous 3-5 joint MNVs), which account for 10.45% of our MNV catalog. In addition, by comprehensive identification of MNVs in 700,134 human samples, we significantly expanded the human MNV landscape.

Functional annotation of MNVs in non-coding regions: While Wang *et al.* and other studies focused on the annotation of MNVs in coding regions, our study extends to the systematic annotation of MNV effects in non-coding regions, showing that 72.47% of pre-miRNA MNVs alter RNA secondary structure, 39.83% of promoter MNVs could create novel TFBS, and 62.60% of 3'UTR regions in protein-coding

genes harbour miRBS-altering MNVs.

4. I recognize one of the key uniqueness of this work compared to existing ones is the functional annotation of non-coding regions, defining TFBS/miRBS creating/disrupting MNVs.

These are really interesting (and the experimental result is beautiful!) but could be further analyzed by somewhat orthogonal aspects. For example,

* For TFBS creating/disrupting MNVs, do we see these to be disrupted for particular class of TFs? Does the MNV rate reflect constraint of TFs and/or the target genes? For example, in Wang *et al*, gained nonsense MNVs were disrupted for constrained genes, which provided an orthogonal evidence that gained nonsense MNVs are more functionally impactful compared to other classes. Could the author provide similar discussion suggesting that TFBS MNVs are indeed functionally impactful?

* For non-coding regions especially, GWAS association itself often times does not imply causality, as there are many variants in linkage (LD), unless statistically fine-mapped with high PIP.

Do the author deem the miRNA MNVs associated in GWAS to be causal? Fine-mapping the associated loci or at least checking the locus zoom plot to see whether the MNV-constituting SNV is the lead variant would be helpful.

Response: Thank you very much for your positive feedback on our non-coding MNV functional annotation work. We sincerely appreciate the insightful suggestion to provide orthogonal evidence supporting the functional impact of TFBS-related MNVs. To address your concerns, we performed three complementary analyses, as detailed below:

(1) MNV rates in TF and TFBS regions.

As shown in Figure R1, we observed that TFBS regions exhibited significantly lower MNV/SNV ratios. Extending this analysis to the coding regions of transcription

factors, we found that TF genes also displayed a significantly reduced MNV/SNV ratio compared to non-TF genes (Wilcoxon rank-sum test, p -value = 3×10^{-10} , Figure R6). From an evolutionary perspective, this pattern indicates that MNVs in both TFBSs and TF genes are under stronger evolutionary constraint, consistent with the critical roles of both *cis*-regulatory elements and TFs in gene regulation.

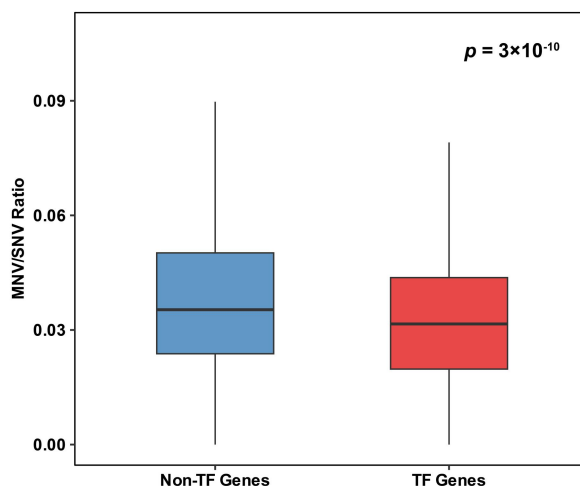


Figure R6. Comparison of MNV/SNV ratios between non-TF genes and TF genes.

(2) Enrichment analysis of MNV-related TFBSs across TF families

Based on TF family annotations from JASPAR, we mapped each TF to its corresponding TF family. To adjust the genomic background differences among TF families, we normalized the MNV-TFBS results by calculating the ratio of predicted gain/loss MNV-TFBSs to all potential TFBSs in each family. The results revealed that MNVs are more likely to impact some specific TF families, such as SAND domain factors, basic helix-span-helix factors, and p53 domain factors (Figure R7). To further explore the possible reasons of this differential enrichment, we analyzed the GC content, which is related to the generation of MNVs, of the binding sequences for each TF family and found a strong positive correlation between GC content and MNV enrichment (Spearman $R = 0.91$, p -value = 3.7×10^{-7} , Figure R8). This correlation suggests that the sequence motif characteristics of TF binding sites may contribute to the differential occurrence of MNVs.

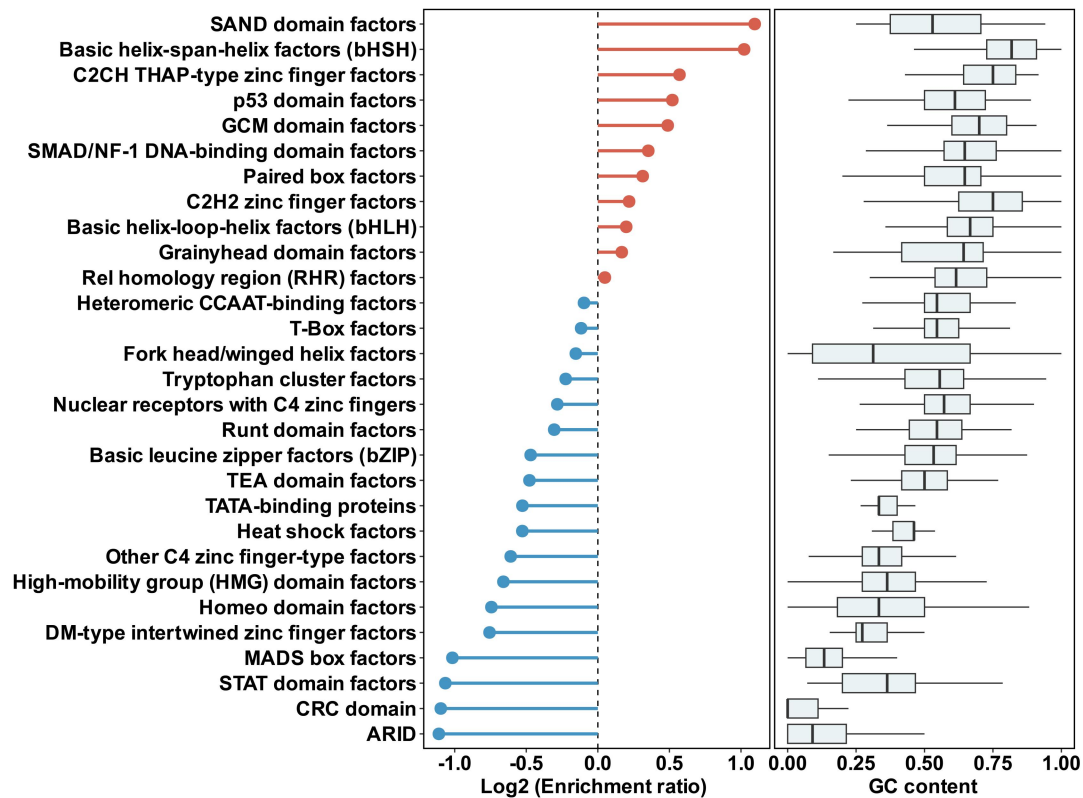


Figure R7. Enrichment of MNV-related TFBSs and the GC content across TF families. The enrichment ratio is defined as the observed rate of predicted gain/loss MNV-TFBSs to all potential TFBSs for each TF family.

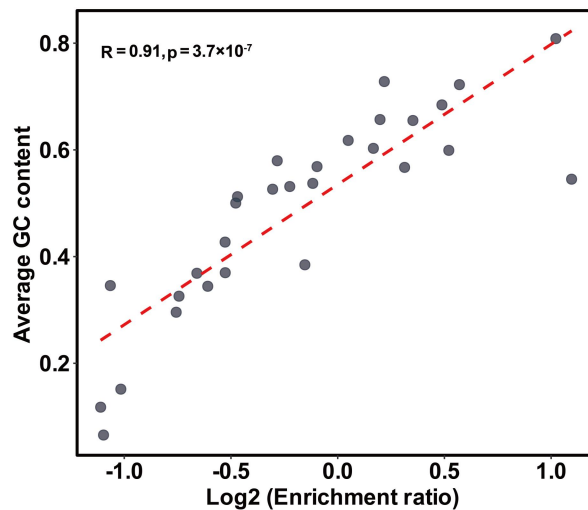


Figure R8. Correlation between the GC content of TFBSs and the log₂ enrichment ratio of TF families.

(3) Analysis of MNV-related TFBSs on functionally constrained genes

To test whether MNVs that alter TFBSs (gain or loss) preferentially avoid the regulatory regions of constrained genes, we performed the LOF intolerance analysis referring to Wang *et al.*'s analysis. We assigned each MNV to its nearest gene and stratified the genes into constrained and non-constrained groups based on gnomAD LOEUF scores (Loss-of-Function Observed/Expected Upper Bound Fraction), a well-established metric for quantifying gene-level functional constraint. We found that the proportion of TFBS-altering MNVs falling in constrained genes is depleted (Figure R9a). In addition, we also observed that the likelihood of MNVs exerting TFBS-altering effects in constrained genes was significantly lower than in non-constrained genes, with an odds ratio of 0.62. (Fisher's exact test, p -value $< 2.2 \times 10^{-16}$; Figure R9b).

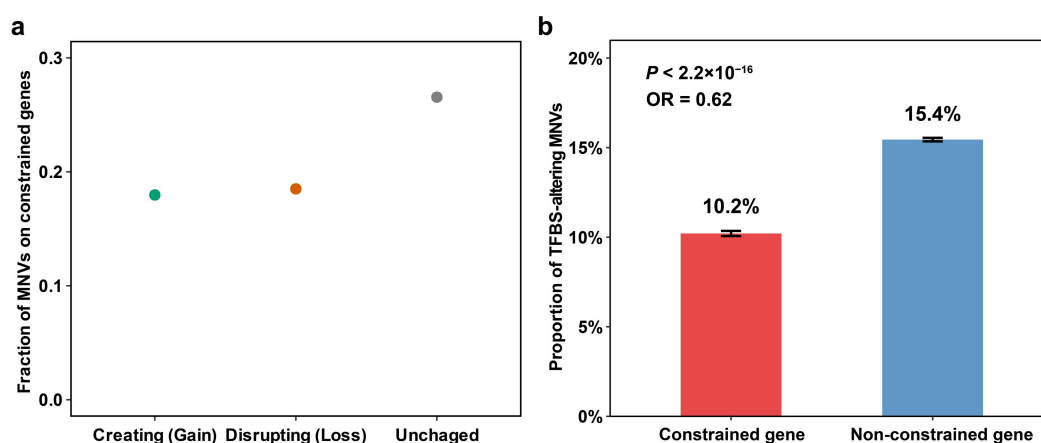


Figure R9. Fraction of each category of MNV found in the constrained genes. (a) Fraction of MNVs (stratified by their TFBS functional categories: Creating/Gain, Disrupting/Loss, Unchanged) located in the regulatory regions of functionally constrained genes. **(b)** Proportion of TFBS-altering MNVs in functionally constrained genes versus non-constrained genes. Statistical significance was evaluated via Fisher's exact test (p -value $< 2.2 \times 10^{-16}$)

Regarding the causality in GWAS results, we fully agree with your perspective that GWAS only demonstrates association rather than causality, particularly in regions with strong linkage disequilibrium (LD). MNVs and their constituent SNVs are

typically in close genomic proximity with inherent relatively strong LD, posing significant challenges for distinguishing true causal variants from tag variants using statistical methods; in this context, definitive causal inference often requires rigorous biological experiments. In our revised manuscript, we have explicitly framed our findings as associations and do not claim that the identified MNVs are causal.

Nevertheless, we want to emphasize the unique value of MNV identification and annotation in GWAS and post-GWAS research in three key aspects: 1) For MNVs not in strong LD with their constituent SNVs, their population frequency distributions may differ from individual SNVs. This can lead to distinct GWAS association signals that would be missed if only SNVs are considered. 2) During post-GWAS candidate variant selection, relying solely on SNV annotation may overlook functionally impactful MNVs—especially those with non-additive effects that are not predictable from individual SNV functions. 3) In candidate genomic regions where MNVs exist as the predominant haplotypic form (due to strong LD), experimental validation should prioritize MNV sequences over individual SNV sequences. This ensures that functional assays reflect the actual genetic context present in the population, rather than artificial single-variant perturbations.

We have integrated these results, along with the Response Figures 7-9 (consolidated into a single new Supplementary Fig. 11), into the revised manuscript. In lines 237-247:

“To further characterize MNV-related TFBSs, we performed an enrichment analysis of MNV-TFBSs across different TF families. After normalizing the ratio of predicted altered MNV-TFBSs to all potential TFBSs in each family, we found that MNVs are more likely to impact some specific TF families (Supplementary Fig. 11a). Analysis of GC content, which was related to the generation of MNVs, in TF binding sequences showed a strong positive correlation with MNV enrichment (Spearman $r = 0.91$, p -value = 3.7×10^{-7} , Supplementary Fig. 11b). Additionally, we characterized the distribution of TFBS-altering MNVs among genes and found that, in functionally

constrained genes, the proportion of TFBS-altering MNVs was lower relative to MNVs with no impact on TFBSs (Supplementary Fig. 11c, d), indicating that TFBS-altering MNVs in functionally constrained genes may also be subject to evolutionary constraint.”

5. Fig 5a is very interesting comparing the delta MFE, but the real key question is to see whether the delta MFE of an MNV is more/less than the sum of delta MFE of the constituent SNV1 and SNV2.

This is analogous to, for example, gained nonsense MNVs where the combined effect is clearly beyond the summation of individual SNVs.

Also, the experimental result the authors provide is a beautiful example where the effect is beyond additive.

Could the author make such comparisons for each miRBS MNVs and discuss MNVs effect beyond/below additive?

Response: Thanks very much for this insightful comment and valuable suggestions.

MNVs are defined as the coexistence of two or more nucleotide variants on the same haplotype in an individual. This definition is broad, encompassing a large number of MNVs with relatively long distances. MNVs with long distances are more likely to exert cumulative effects of their constituent SNVs (i.e. additive effects). Therefore, consistent with the approach adopted by Wang *et al.*, we focused on MNVs with a genomic distance of ≤ 10 bp, where MNVs are more likely to alter codons or regulatory motifs in ways distinct from their constituent SNVs, leading to functional effects that differ from the sum of their individual constituent SNVs (i.e., non-additive effects). For example, in the coding region, we focused on finding MNVs that may have different effects on amino acid coding compared to their constituent SNVs. In the non-coding region, we focused on finding MNVs that may affect functional elements, such as miRBS and TFBS.

In Figure 5a, our initial analysis focused on comparing Δ MFE (delta Minimum Free

Energy) differences in the secondary structure energy between MNVs and SNVs. Following your suggestion, we extended this analysis by computing the sum of Δ MFEs of the constituent SNVs, and then compared it with the corresponding MNV Δ MFE. The results revealed that 52.16% of MNVs exhibited Δ MFE values deviating from the sum of their constituent SNVs' Δ MFE (Figure R10), indicating non-additive effects on ncRNA secondary structure stability.

In addition, nucleotide sequences are the fundamental determinants of biological function. Our experiment results also show that an MNV located in the miRNA seed region can induce widespread alterations in target genes. Similarly, sequence changes in the miRNA binding sites within gene 3'UTRs can also lead to non-additive changes in miRNA-mRNA interactions.

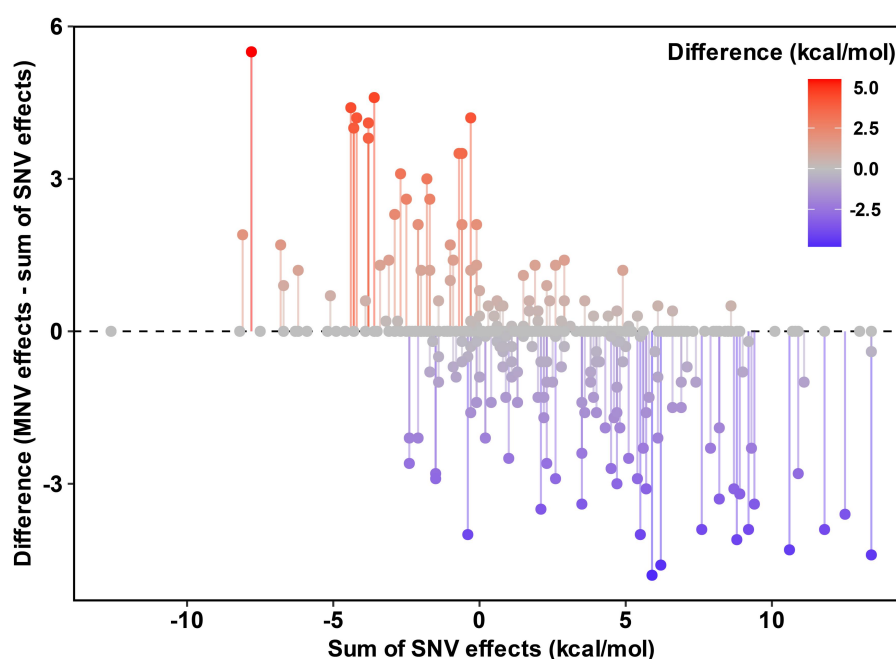


Figure R10. Divergence in Δ MFE between MNV effects and cumulative SNV effects

In theory, accurately distinguishing between additive and non-additive MNV effects is challenging without direct experimental validation. However, MNVs highlighted in our manuscript and database, including unique amino acid changes in coding regions, distinct impacts on secondary structure energy, target gene changes in miRNA seed

regions, and altered miRNA and TF binding in non-coding regions, are more likely to be non-additive.

We have integrated these results into the revised manuscript in lines 266-275:

“Our results showed that 396 MNVs were present in pre-miRNAs, and 287 (72.47%) of which could alter the secondary structure of pre-miRNAs ($|\Delta \text{Minimum Free Energy } (\Delta \text{MFE})| > 1 \text{ kcal/mol}$), and MNVs exhibited significantly greater $|\Delta \text{MFE}|$ than individual SNVs (Fig. 5a). For example, compared with wild-type and SNV-mutated pre-miRNAs, the secondary structure of MNV-mutated hsa-mir-1307 exhibited significant changes (Fig. 5b). Additionally, we calculated the sum of ΔMFEs of the constituent SNVs for each MNV, and compared this sum with the corresponding MNV ΔMFE . The results revealed that 52.16% of MNVs exhibited ΔMFE values deviating from the sum of their constituent SNVs' ΔMFE (Supplementary Fig. 12), indicating non-additive effects on ncRNA secondary structure stability.”

Minor comments:

- Line 85 “high” and “low” MNVs should be clearly defined.

Response: Thank you for pointing out this issue. In the context of this study, the terms "high" and "low" are relative concepts. For example: compared to the 2-joint MNVs, the 3-joint MNV belongs to the "high" category, while the 2-joint MNVs fall into the "low" category.

We have revised the manuscript accordingly to clarify these definitions. In lines 97-102:

“Thus, we developed MNVAnno (See Supplementary Materials for details), a toolbox for the identification and annotation of all kinds of MNVs (Fig. 1c). MNVAnno adopts a sub-joint MNV merging strategy, thereby avoiding the possibility of

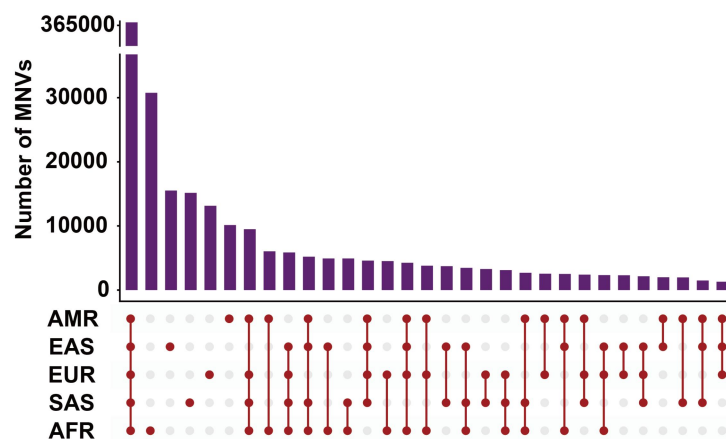
mislabeling higher joint MNVs as lower joint MNVs (e.g., a 3-joint MNV is misidentified as multiple 2-joint MNVs), and provides more comprehensive annotation in both coding and non-coding regions (Table 1).”

- Line 106 and other places: GnomAD should be gnomAD.

Response: Thank you for this correction. We have revised all the "GnomAD" to "gnomAD" throughout the manuscript.

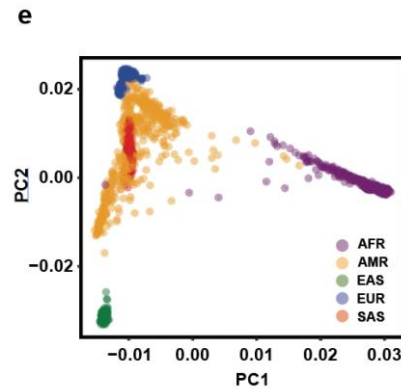
- Fig 2d and several other figures: Venn diagram with more than 3 entries is very hard to visualize. Could the authors use upsetr plot instead?

Response: Thank you for the valuable suggestion. We agree that upset plots provide a clearer and more effective visualization for sets with more than three entries. Following your recommendation, we have replaced the Venn diagrams in Fig. 2d with upset plots in the revised manuscript.



- Fig. 2e: x and y axis should be PC1 and PC2 rather than "PCA" 1 and 2? (A stands for "Analysis" and should not be needed here for the legend).

Response: Thank you for pointing out this issue. We have updated Fig. 2e accordingly, changing the axis labels from "PCA 1" and "PCA 2" to "PC1" and "PC2" in the revised manuscript.



- Line 120-127: The biological connection here is a bit superficial. Is the number "12 SNVs" more than expected by chance? I would not mind deleting this entire discussion, or providing some additional details.

Response: Thank you for this comment. We agree that the biological connection established by KEGG enrichment was somewhat simplistic and superficial. However, as KEGG enrichment is commonly used analytical approach in related studies and another reviewer thought this part of interest, we have provided more details about this.

We extracted the genes significantly enriched in the Cardiomyopathy-related pathway and calculated the F_{ST} values of MNVs and SNVs in these genes, respectively. The results showed that the F_{ST} values of MNVs are significantly higher than those of SNVs (p -value = 0.001) (Figure R11), suggesting that MNVs may exert some effects on cardiomyopathy in African populations.

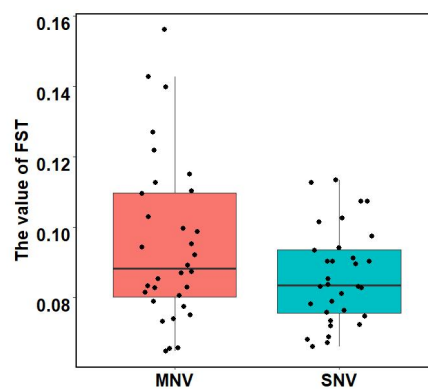


Figure R11. The F_{ST} values of MNVs and SNVs in cardiomyopathy-related genes

We have incorporated these results and the corresponding Response figure into the revised manuscript. The relevant content can be found in lines 149–159 of the updated text.

“We counted the number of MNVs with high F_{ST} values ($F_{ST} \geq 0.15$) across populations and found that the African population had the largest number of such MNVs (Fig. 2g). The KEGG analysis showed that genes harboring these MNVs were significantly enriched in the cardiomyopathy-related pathways (Fig. 2h). Moreover, we extracted the genes significantly enriched in the cardiomyopathy-related pathway and calculated the F_{ST} values of MNVs and SNVs in these genes respectively. The results showed that the F_{ST} values of MNVs were significantly higher than those of SNVs (p -value = 0.001) (Supplementary Fig. 7). Several studies have reported that the incidence of cardiomyopathy varies in different populations, and dilated cardiomyopathy is a major cause of heart failure across Africa²³. Collectively, these findings may provide novel insights for the study of dilated cardiomyopathy in African populations.”

- Line 161-163: This MNV density enrichment could probably be explained simply by SNV mutation rate differences?

Response: Thank you for raising this thoughtful question regarding the potential influence of SNV mutation rate differences on the observed MNV density enrichment.

In the response to your first major comment, we observed that MNV density enrichment in our analysis cannot be solely attributed to differences in the background SNV mutation rate. As illustrated in Figure R1 a and b, after normalizing for SNV density (i.e., adjusting for SNV ratios), several highly conserved genomic regions (exons, splice sites, conserved elements, miRBS, and TFBS) exhibited significantly lower MNV/SNVR ratios. This pattern suggests stronger evolutionary constraints against MNVs in these functionally critical regions, implying that MNVs are likely

more deleterious than individual SNVs from an evolutionary perspective. Please refer to our response to the first comment for more details.

- Fig. 5h experiment: Why is the NC group somewhere in between WT and MNV? Is this expected?

Thank you for your careful observation. This experiment was designed to validate whether MNVs can exert distinct functional effects in biological processes. The prediction results indicate that only the MNV mutant of hsa-miR-585-3p can bind to the 3'UTR of the *PLAU* gene, thereby inhibiting its expression. This outcome is consistent with the expectations. Regarding the phenomenon that the WT group (wild-type group) shows higher expression compared to the NC group (negative control group), it is speculated that, apart from the currently known pathways of action, hsa-miR-585-3p may also indirectly affect the expression of the *PLAU* gene through other unknown signaling pathways or by regulating other related genes. However, the specific regulatory mechanisms require further investigation.

- Line 246-247: Is this more than expected by chance? In other words, are these MNVs enriched in oncogenes, compared to random MNVs?

Response: Thank you for this comment. To address this question, we downloaded oncogene lists from NCG7.1 or intOGen database, then compared the proportion of cancer GWAS-identified MNVs located in oncogenes with that in non-oncogenes. Statistical analysis revealed a significant enrichment of these MNVs in known oncogenes (Fisher's exact test, p -value = 0.038), indicating a non-random distribution.

- line 250: Should clarify that $p < 0.05$ is “nominally” significant, or apply multiple test correction.

Response: Thank you for this comment. In the revised manuscript, we have clarified this on lines 325-327:

“Subsequently, we evaluated the impacts of these MNVs on cancer prognosis using three models, including additive, dominant, and recessive models, and identified 46 associated MNVs with nominal significance (p -value < 0.05)”

The webpage is a nice browser but I don't see the source code (maybe I am missing it)

Response: Thanks very much for your positive feedback about our webpage. The MNVAnno source code can be accessed via the following link:

<https://gonglab.hzau.edu.cn/MNVAnno/download/download.html>

[Figure Redacted]

Reviewer #2

The manuscript “Genome-wide identification and annotation of multi-nucleotide variants in human reveal widespread functional differences compared to single nucleotide variants” presents a large-scale analysis of multi-nucleotide variants (MNVs) across over 700,000 human samples. The authors develop MNVAnno, a new tool for identifying and annotating MNVs, and compile the largest human MNV catalog to date. They demonstrate that MNVs can have distinct functional consequences compared to single nucleotide variants (SNVs), affecting both coding and non-coding regions, and identify potential disease associations, including with several cancers. The work provides valuable resources and an elegant analysis, but certain methodological and interpretative aspects require clarification and strengthening.

Response: Thank you very much for the time and thoughtful comments you devoted to reviewing our manuscript. We greatly appreciate your positive assessment of our work and its value to the research community. We have carefully addressed all raised concerns with additional analyses, and revised the manuscript accordingly. Please see the point-to-point response below.

Comments

1. Definition and complexity of MNVs

In the introduction, the manuscript refers to “complex MNVs” but does not define what qualifies as complex. Additionally, a more detailed explanation of the challenges in MNV annotation is needed. This would set the context for why existing tools struggle and what precisely MNVAnno improves.

Response: We greatly appreciate your constructive comment and the insightful suggestion to elaborate on these key issues in detail.

For the complexity of MNVs, MNVs are defined as the coexistence of two or more

nucleotide variants on the same haplotype within an individual. As shown in Figure R1, their combinations are diverse and complex, including 2-joint and multi-joint MNVs (≥ 3 -variants) spanning coding and non-coding regions, as well as adjacent continuous and non-continuous variants. Traditional identification tools do not comprehensively consider these complex MNV situations. For example, VarDict only identifies continuous multi-joint MNVs, MACARON only supports MNVs within a single codon, and Wang's tool is suitable for 2-joint MNVs (Table R1).

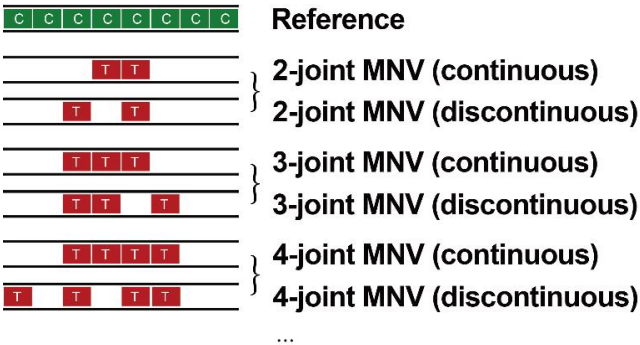


Figure R1. Illustrative examples of the different MNV types

For the challenges of MNV identification, existing tools face significant limitations in addressing MNVs, which are primarily manifested in three key aspects: **1) Narrow identification scope.** MNVs are defined as coexisting variants on the same haplotype. Their combinations are diverse and complex, including 2-joint and multi-joint MNVs (≥ 3 -variants, accounting for ~10% of our results) spanning coding and non-coding regions, as well as adjacent continuous and non-continuous variants (~80% of all MNVs). However, most current methodologies only identify specific types of MNVs. For instance, VarDict only identifies continuous multi-joint MNVs, MACARON is restricted to MNVs within a single codon, BCFtools focuses on exonic MNVs in intron-free General Feature Format (GFF) regions, and Wang's tool is designed for 2-joint MNVs. **2) Inadequate annotation.** The majority of MNVs are located in non-coding regions; however, most existing tools only annotate coding 2-joint MNVs, overlooking non-coding MNV functions (e.g., MNVs in ncRNAs, promoters and regulatory regions). **3) High time and memory consumption.** Tools like MAC require > 10 -fold more memory and longer runtime than MNVAnno, limiting their

applicability to large-cohort analyses. In summary, these limitations stem from the interplay of MNV's inherent complexity and tool design constraints, underscoring the urgent need for MNV identification and annotation.

Table R1. Comparison of the application scope of six tools

Tools	Input data type	Support multiple samples	Identification scope	Supported MNV types	Annotation scope
VarDict	BAM	No	Genome	Only support continuous multi-joint MNVs	Not support
MAC	SNP (VCF) BAM	No	Genome	Support multi-joint MNVs	Function annotation in coding regions
MACARON	VCF	No	Codon	Support MNVs within a single codon	Function annotation in coding regions
BCFtools	Phased VCF	Yes	Intron-free (GFF)	Support multi-joint MNVs	Not support
Wang's tool	Phased VCF	Yes	Genome	Support 2-joint MNVs	Function annotation in coding regions and position annotation in other regions
MNVAnno	Phased VCF	Yes	Genome	Support multi-joint MNVs	Function annotation in coding regions and non-coding regions (e.g. ncRNA, 3'UTR, TFBS)

For the improvement of MNVAnno, it addresses these critical limitations through three key advantages: **1) Comprehensive MNV identification.** Systematically detects complex MNV types (multi-joint, and non-continuous) across the entire genome. **2) Genome-wide functional annotation.** Integrates functional interpretation for both coding and non-coding regions (e.g., transcription factor binding sites [TFBS], miRNA binding sites [miRBS], ncRNA secondary structure). **3) Computational**

efficiency. MNVAnno shows extremely low runtime and the lowest maximum memory usage among comparable tools.

In the revised manuscript, we have updated the Table 1 and incorporated relevant content into lines 91–99.

“By definition, besides 2-joint continuous MNVs, there are also various other types of MNVs, including multi-joint MNVs (≥ 3 variants) as well as adjacent continuous and non-continuous variants spanning coding and non-coding regions (Supplementary Fig. 1). Current tools primarily focus on continuous MNVs within coding regions. For example, VarDict only identifies continuous multi-joint MNVs¹⁶, MACARON is limited to MNVs within a single codon¹⁷, BCFtools focuses on exonic MNVs in intron-free General Feature Format (GFF) regions¹⁸, and Wang’s tool is primarily suitable for 2-joint MNVs⁹ (Table 1). Thus, we developed MNVAnno (See Supplementary Materials for details), a toolbox for the identification and annotation of all kinds of MNVs (Fig. 1c). ”

2. Sources of false positives and false negatives

The introduction or Methods should discuss in more depth the sources of error in MNV annotation. What leads to false positives or false negatives? Is it related to phasing quality only, or other factors are important?

Response: We thank the reviewer for this valuable comment. To address it, we further carefully analyzed the comparison results, and found that the sources of false positives and false negatives for various methods stem primarily from three aspects: supported MNV types and identification scope, input data requirements, and sub-joint variant merging capability.

(1) Supported MNV types and identification scope. The most prominent discrepancies across tools arise from their supported MNV types and genomic coverage. VarDict only detects continuous MNVs; Wang's tool mainly supports

2-joint MNVs and a few 3-joint MNVs within coding regions; MACARON is restricted to single codons; BCFtools focuses on exonic MNVs in intron-free General Feature Format (GFF) regions. MAC handles multi-joint MNVs genome-wide but identifies far fewer than others. In contrast, MNVAnno can handle both continuous and discontinuous ones (~80% of all MNVs), and supports 3-5 joint MNVs (~10% of our results) across the entire genome.

(2) Input data requirements

BAM-based tools (VarDict): VarDict relies on raw sequencing reads (BAM files) to call MNVs by detecting adjacent variants in the same read. However, it cannot verify haplotype phase (cis/trans status) of variants. This leads to false positives (e.g., incorrect read mapping and haplotype phase error) and false negatives (cannot identify non-continuous MNVs).

Phased VCF-based tools (MNVAnno, Wang's tool, BCFtools): These tools use phased VCFs to avoid phase-related errors. However, their performance depends on phasing quality—rare errors (mean error rate $< 0.005^{11}$) still introduce minor inaccuracies.

MAC requires both a SNP list from VCF files and BAM files for its operation. It infers haplotypes from reads spanning multiple SNVs, defines connected components as “Block of Mutation (BM)”, and then identifies MNVs based on BMs. MAC was published in 2015 and identifies far fewer MNVs than other tools. However, the authors have not provided specific details regarding its identification criteria or filtering thresholds, which makes it difficult to pinpoint the exact reasons for its discrepancies compared to other methods.

(3) Sub-joint MNV merging capability

MNVAnno employs a "high-joint priority" identification strategy (Figure R2). For example, if three adjacent or closely spaced variants are detected on the same haplotype, MNVAnno directly classifies them as a single 3-joint MNV instead of splitting them into multiple 2-joint sub-sets (e.g., variants at positions 100, 102, 103

are annotated as one 3-joint MNV, not three separate 2-joint MNVs: 100+102, 102+103 and 100+103). This merging step prevents misclassification of sub-joint variants as independent MNVs and preserves the biological authenticity. In contrast, other tools lack this merging step, leading to redundant calls of overlapping sub-joint MNVs that reduce precision and recall.

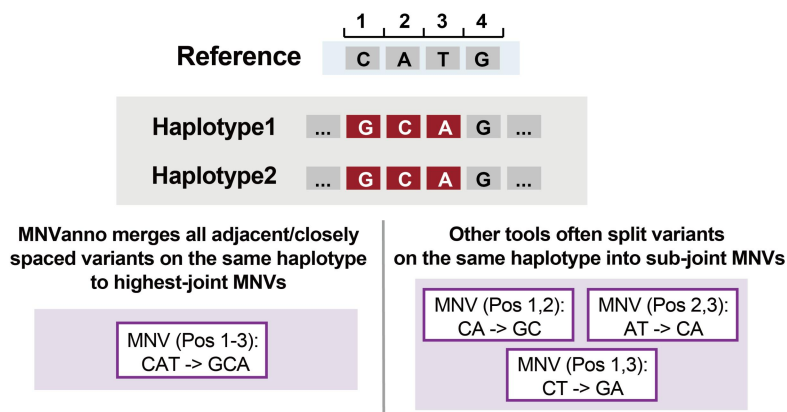


Figure R2. Sub-joint MNV merging strategy

However, this strategy introduces a minor trade-off in comparative evaluations across tools, and it is the main reason MNVAnno fails to achieve 100% precision/recall in gold-standard dataset comparisons. MNVAnno's merging strategy may lead to occasional false positives and false negatives due to rare phasing errors. For instance, if a 3-joint MNV identified by MNVAnno includes one variant with phasing errors (e.g., an artifact SNV), the entire 3-joint MNV results in a false positive classification. Meanwhile, this 3-joint MNV is split into multiple sub-joint MNVs by other tools (e.g., Wang's tool), and those sub-joints have sufficient read support, MNVAnno may be marked with false negatives for failing to call those sub-joints independently. We carefully examined the comparison results of MNVAnno. There were 391 false positives and 1,436 false negatives, and most of them were caused by high-joint MNV combination. Nevertheless, we still maintain that reporting the observed haplotype as a single MNV remains biologically more faithful than enumerating redundant sub-sets, and this principle aligns with our emphasis on analyzing MNV rather than individual constituent SNVs.

In the revised manuscript, we have updated the Table 1 and incorporated relevant content into lines 91–102 and lines 116-121.

“By definition, besides 2-joint continuous MNVs, there are also various other types of MNVs, including multi-joint MNVs (≥ 3 variants) as well as adjacent continuous and non-continuous variants spanning coding and non-coding regions (Supplementary Fig. 1). Current tools primarily focus on continuous MNVs within coding regions. For example, VarDict only identifies continuous multi-joint MNVs¹⁶, MACARON is limited to MNVs within a single codon¹⁷, BCFtools focuses on exonic MNVs in intron-free General Feature Format (GFF) regions¹⁸, and Wang’s tool is primarily suitable for 2-joint MNVs⁹ (Table 1). Thus, we developed MNVAnno (See Supplementary Materials for details), a toolbox for the identification and annotation of all kinds of MNVs (Fig. 1c). MNVAnno adopts a sub-joint MNV merging strategy, thereby avoiding the possibility of mislabeling higher joint MNVs as lower joint MNVs (e.g., a 3-joint MNV is misidentified as multiple 2-joint MNVs), and provides more comprehensive annotation in both coding and non-coding regions (Table 1)”

“The discrepancies among tools mainly stem from three aspects: (i) supported MNV types (e.g., VarDict only identifies continuous MNVs, while MNVAnno covers continuous and non-continuous multi-joint MNVs); (ii) input data requirements (which may introduce alignment errors and phasing errors); and (iii) merging strategy for high-joint MNVs (MNVAnno adopts “high-joint priority” strategy to prevent splitting 3-joint MNVs into multiple 2-joint ones, a feature not available in other tools) (See Supplementary Materials for details).”

3. Normalization for sample size and mutation rate (Figure 4a)

The authors compare MNV densities across datasets and cohorts, but do not correct for differences in sample size or mutation rate. Models such as the Roulette or Chen *et al.* (Nature 2024) mutation rate model could be used for normalization. Without this,

density differences between cohorts reflect mainly differences in sample size, which is not an interesting feature. More importantly, the density difference between genomic categories can be caused by mutation rate differences.

Response: Thank you very much for this insightful and constructive comment, which helps us better understand MNV densities across different genomic regions.

To directly address this question, we performed the following analyses.

To eliminate the influence of sample size or mutation rate, we calculated the MNV/SNV Ratio (MNV density normalized by SNV density) for all functional genomic regions, with the genome-wide average ratio (the number of all MNVs / the number of all SNVs) set as the baseline, using whole-genome sequencing (WGS) data from 1000G and GTEx datasets, respectively. As shown in Figure R3a-b, we observed that in most regions (e.g., intron, lncRNA, miRNA, 3'UTR), the MNV/SNV ratios are around the baseline, indicating that MNV occurrence in these regions largely parallels the SNV mutation rate. Highly conserved regions (e.g., exon, splice, conserved element, miRBS, and TFBS) exhibited significantly lower MNV/SNV ratios. This reflects an evolutionary constraint against MNVs in these functionally critical regions, indicating that MNVs are more deleterious than individual SNVs from the evolutionary aspects. Interestingly, after normalization, we observed that the 5'UTR region showed a relative enrichment of MNVs. Prior studies have indicated that the 5'UTR region is rich in CpG dinucleotides ^{1,2}, which are more likely to induce the generation of MNVs ³.

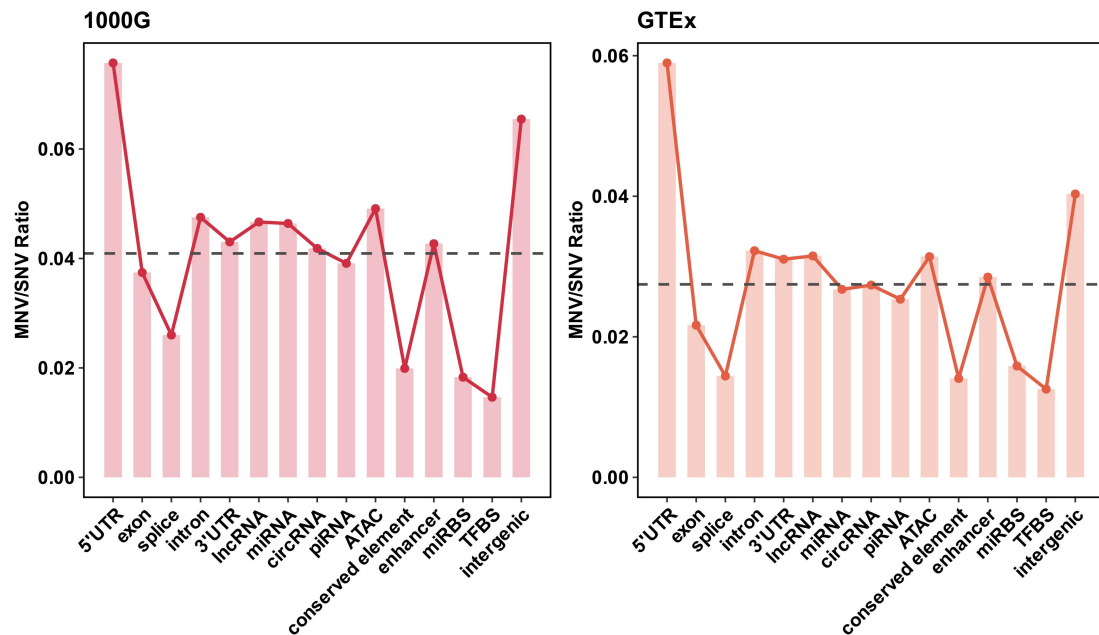


Figure R3. The distribution of the MNV/SNV ratios in different regions of 1000 Genomes and GTEx. (a,b) MNV/SNV ratios across annotated genomic regions in the 1000 Genomes Project (1000G; a) and GTEx (b). The dashed horizontal line represents the genome-wide background (baseline) MNV/SNV ratio.

We have integrated these results and Response figure into the revised manuscript. In the revised manuscript, we have added the following sentences in lines 195-208.

“After normalizing the genomic length, a similar density distribution was observed across whole-genome sequencing datasets (Fig. 4a). Moreover, to correct for regional differences in SNV mutation rate, we calculated the MNV/SNV ratio (MNV density normalized by SNV density) for all functional genomic regions in WGS datasets. We observed that in most regions (e.g., intron, lncRNA, miRNA, 3'UTR), the MNV/SNV ratios were around the baseline, indicating that MNV occurrence in these regions largely paralleled the SNV mutation rate (Supplementary Fig. 9). Highly conserved regions (exon, splice region, conserved element, miRBS, and TFBS) exhibited significantly lower MNV/SNV ratios. This reflects an evolutionary constraint against MNVs in these functionally critical regions, indicating that MNVs are more deleterious than individual SNVs from the evolutionary perspective. Notably, after normalization, we observed that the 5'UTR region exhibited a relative enrichment of

MNVs (Supplementary Fig. 9). Previous studies have indicated that the 5'UTR region is rich in CpG dinucleotides^{29,30}, which are more likely to promote the generation of MNVs⁹.”

4. Cardiomyopathy association (lines 120-126)

The enrichment of MNVs in cardiomyopathy pathways in African populations is interesting but should be compared to the same analysis using only SNVs. It is unclear whether this signal reflects truly MNV-specific effects or the aggregation of constituent SNVs already known from GWAS.

Response: We comprehensively examined the SNVs with high FST values. The results showed that these high-FST SNVs were almost widely distributed across all genes. Specifically, there were 13,211 genes containing three high-FST SNVs, and 9,667 genes containing ten high-FST SNVs. Given this situation, it was not appropriate to conduct an enrichment analysis on all high-FST SNVs.

Therefore, we extracted the genes significantly enriched in the Cardiomyopathy-related pathway and calculated the FST values of MNVs and SNVs in these genes, respectively. The results showed that the FST values of MNVs are significantly higher than those of SNVs (p value = 0.001) (Figure R4). The results suggest that MNVs may exert a more significant MNV-specific effect on the complex phenotype of cardiomyopathy in African populations.

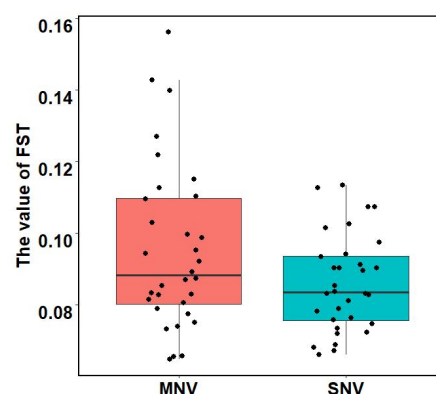


Figure R4. The FST values of MNVs and SNVs in cardiomyopathy-related genes

The revised manuscript now contains these results and the Response figure, with corresponding adjustments made in lines 149–158.

“We counted the number of MNVs with high F_{ST} values ($F_{ST} \geq 0.15$) across populations and found that the African population had the largest number of such MNVs (Fig. 2g). The KEGG analysis showed that genes harboring these MNVs were significantly enriched in the cardiomyopathy-related pathways (Fig. 2h). Moreover, we extracted the genes significantly enriched in the cardiomyopathy-related pathway and calculated the F_{ST} values of MNVs and SNVs in these genes respectively. The results showed that the F_{ST} values of MNVs were significantly higher than those of SNVs (p -value = 0.001) (Supplementary Fig. 7). Several studies have reported that the incidence of cardiomyopathy varies in different populations, and dilated cardiomyopathy is a major cause of heart failure across Africa²³.”

5. Additivity of functional effects (Figure 5a)

Figure 5a shows that MNVs produce larger changes in RNA secondary structure ($|\Delta MFE|$) than SNVs, but the crucial question is whether these effects are non-additive. The authors should compare the MNV effect to the sum of the effects of its constituent SNVs to demonstrate epistatic or synergistic interactions—this is central to the biological rationale for analyzing MNVs as combined events.

Response: Thank you very much for this insightful comment and valuable suggestion.

To address this, we calculated the sum of ΔMFE s of the constituent SNVs for each MNV, then comparing it with the corresponding MNV ΔMFE . The results revealed that 52.16% of MNVs exhibited ΔMFE values deviating from the sum of their constituent SNVs' ΔMFE (Figure R5), indicating non-additive effects on ncRNA secondary structure stability.

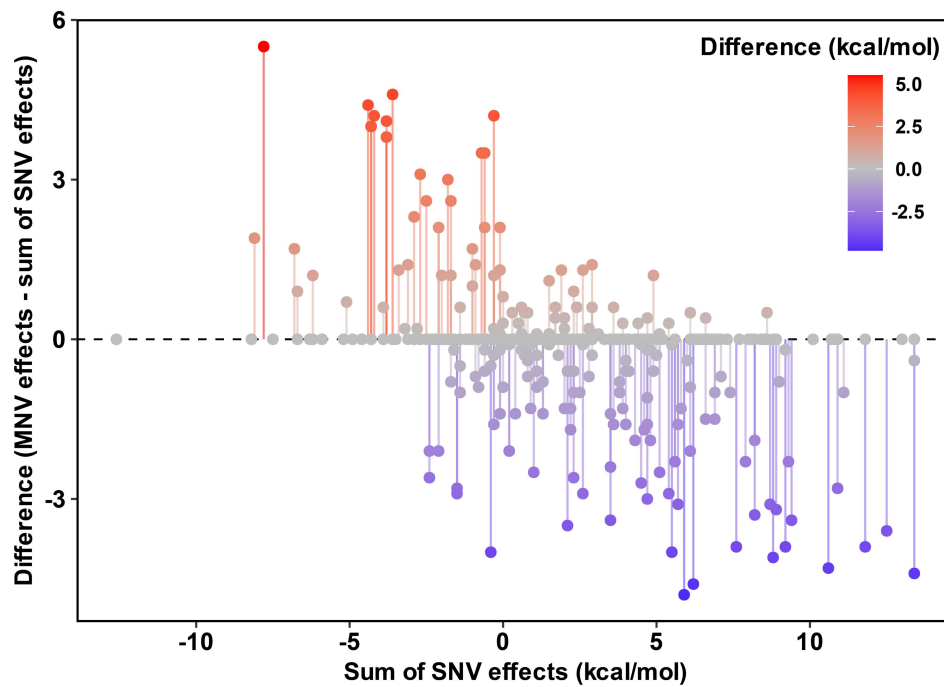


Figure R5. Divergence in Δ MFE between MNV effects and cumulative SNV effects

We have integrated these results into the revised manuscript in lines 266-275:

“Our results showed that 396 MNVs were present in pre-miRNAs, and 287 (72.47%) of which could alter the secondary structure of pre-miRNAs ($|\Delta$ Minimum Free Energy (Δ MFE) | > 1 kcal/mol), and MNVs exhibited significantly greater $|\Delta$ MFE| than individual SNVs (Fig. 5a). For example, compared with wild-type and SNV-mutated pre-miRNAs, the secondary structure of MNV-mutated hsa-mir-1307 exhibited significant changes (Fig. 5b). Additionally, we calculated the sum of Δ MFEs of the constituent SNVs for each MNV, and compared this sum with the corresponding MNV Δ MFE. The results revealed that 52.16% of MNVs exhibited Δ MFE values deviating from the sum of their constituent SNVs' Δ MFE (Supplementary Fig. 12), indicating non-additive effects on ncRNA secondary structure stability.”

6. GWAS significance threshold

The use of $p < 5 \times 10^{-6}$ as a genome-wide threshold (line 246) can be too permissive. The standard threshold for GWAS is 5×10^{-8} , or alternatively, an appropriate Bonferroni or FDR correction should be applied based on the number of MNVs tested. Without this correction, the reported 267 associations may include false positives.

Response: Thank you very much for this comment. Following your suggestion, we used the false discovery rate (FDR) to correct the p values, and finally identified 232 MNVs that were significantly associated with cancer susceptibility (FDR < 0.05).

In the revised manuscript, we have revised the relevant content in lines 318-322:

“MNVs, as an important class of genetic variants, are expected to contribute to extra phenotype diversity. The occurrence and development of cancer are a complex biological process that is affected by both genetics and environments⁴⁸. Here, we performed GWAS analyses on 11 cancers (sample size > 1000) (Supplementary Fig. 15) and obtained 233 MNVs significantly associated with cancer susceptibility (FDR < 0.05) (Supplementary Table 2).”

References

1. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* **103**, 1412–1417 (2006).
2. Li, J. *et al.* CpG hypomethylation at proximal promoter and 5'UTR along with IL6 signaling loop associates with MYD88 upregulation in epithelial ovarian cancer. *Sci Rep* **14**, 30945 (2024).
3. Wang, Q. *et al.* Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. *Nat Commun* **11**, 2539 (2020).
4. Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304 (2000).
5. Francioli, L. C. *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet* **47**, 822–826 (2015).
6. Xue, Y. *et al.* Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Curr Biol* **19**, 1453–1457 (2009).
7. Harris, K. & Nielsen, R. Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome Res* **24**, 1445–1454 (2014).
8. Zhong, X. *et al.* The fidelity of DNA synthesis by yeast DNA polymerase zeta alone and with accessory proteins. *Nucleic Acids Res* **34**, 4731–4742 (2006).
9. Sakamoto, A. N. *et al.* Mutator alleles of yeast DNA polymerase zeta. *DNA Repair* **6**, 1829–1838 (2007).
10. Stone, J. E., Lujan, S. A., Kunkel, T. A. & Kunkel, T. A. DNA polymerase zeta generates clustered mutations during bypass of endogenous DNA lesions in *Saccharomyces cerevisiae*. *Environ Mol Mutagen* **53**, 777–786 (2012).
11. Stahl, K., Gola, D. & König, I. R. Assessment of Imputation Quality: Comparison of Phasing and Imputation Algorithms in Real Data. *Front Genet* **12**, 724037 (2021).

Point-by-point response to the reviewers' comments

Reviewer #1:

I truly appreciate the authors' effort addressing my questions. Most of my original questions are well addressed. I hope the following three remaining points could be further addressed before publication. Thank you so much for the authors' effort in advance.

Response: Thank you very much for your positive feedback on our previous revision. We sincerely appreciate the constructive and insightful suggestions you provided in this second round of review, which have helped improve our study. Below, we provide point-by-point responses to the three remaining questions.

1) Re-visit the interpretation of (MNV/SNV) ratio:

I think the unit of MNV/SNV ratio is a bit tricky - for one-step event, normalization by SNV rate is fine, but strictly speaking, for two-step event, MNV is a combination of SNVs, so the normalization should be dividing by the square of SNV rate (i.e. SNV rate^2 , sorry that I did not clarify this in my original comment. MNV rate should also reflect how uneven the SNVs are distributed, but that is the kind of complexity beyond our analysis.). Probably that reflects the fact that the sharp curve of Type “SNV event” versus relatively flat curve of other one-step type events in Figure R1.

Probably due to this fact, the MNV/SNV ratio might not fully reflect conservation. For example,

- exons and splice regions are in general more conserved than TFBS, but the MNV/SNV ratio does not reflect that.
- 5'UTR should be relatively conserved but as the authors mentioned, we don't see depletion of MNV/SNV ratio because of differences of CpG density which makes the SNV rate different.

The 5'UTR observation makes sense since, let the global SNV rate be p and SNV rate

in a particular functional category be $c \cdot p$. For a two step MNV event, ignoring the positional dependencies for simplicity, MNV rate would scale with $(c \cdot p)^2$, so (MNV/SNV) would be $(c \cdot p)^2 / (c \cdot p) = c \cdot p$. This means that if the mutation rate itself is high in the category (i.e. $c > 1$), the (MNV/SNV) is also going to be high, for two-step MNV events. This could probably explain the fact that 5'UTR, rich in CpG methylation, has high (MNV/SNV) . This also seems to match Fig. 2c, where visually $y = x^2$ could be a better fit.

Therefore, the logic that “low $\text{MNV}/\text{SNV} \Rightarrow$ stronger evolutionary constraint specifically on MNV” may not hold true. I think we need to be careful about the description in the newly added sentences: “This reflects an evolutionary constraint against MNVs in these functionally critical regions, indicating that MNVs are more deleterious than individual SNVs from the evolutionary perspective” - this could be true evolutionary constraint OR just the limitation of the way we normalize the MNV rate.

Given these facts, here are two additional recommendations:

1. Look at the $\text{MNV}/(\text{SNV}^2)$ ratio in addition to the current MNV/SNV ratio, and check which parts of the discussion still hold true.
2. Lower the tone of the sentence above. Instead of mentioning “this reflects an evolutionary constraint”, the authors could write something like “this *could* reflect..., *although it could simply reflect the complex relationship between SNV rate vs MNV rate [ref. 22]*”.

If the authors think 1. is beyond the scope of the paper, just 2. alone is fine.

Response: Thank you very much for this insightful comment on the mathematical relationship between SNV and MNV rates. We agree that for MNVs arising from a two-step process (independent mutations at adjacent sites), the expected MNV rate may scale with the square of the local SNV rate. Thus, we further calculated the MNV/SNV^2 ratio across different functional genomic categories. From the results

(Figure R1), we observed that conserved regions (e.g., exon, splice, conserved element, miRBS, and TFBS) still exhibited relatively lower ratios than other regions (Figure R1a and 1b). Nonetheless, we fully agree that the relationship between SNV rate and MNV rate is complex, and caution is needed in its description. Therefore, we have followed your second recommendation to moderate our claims accordingly.

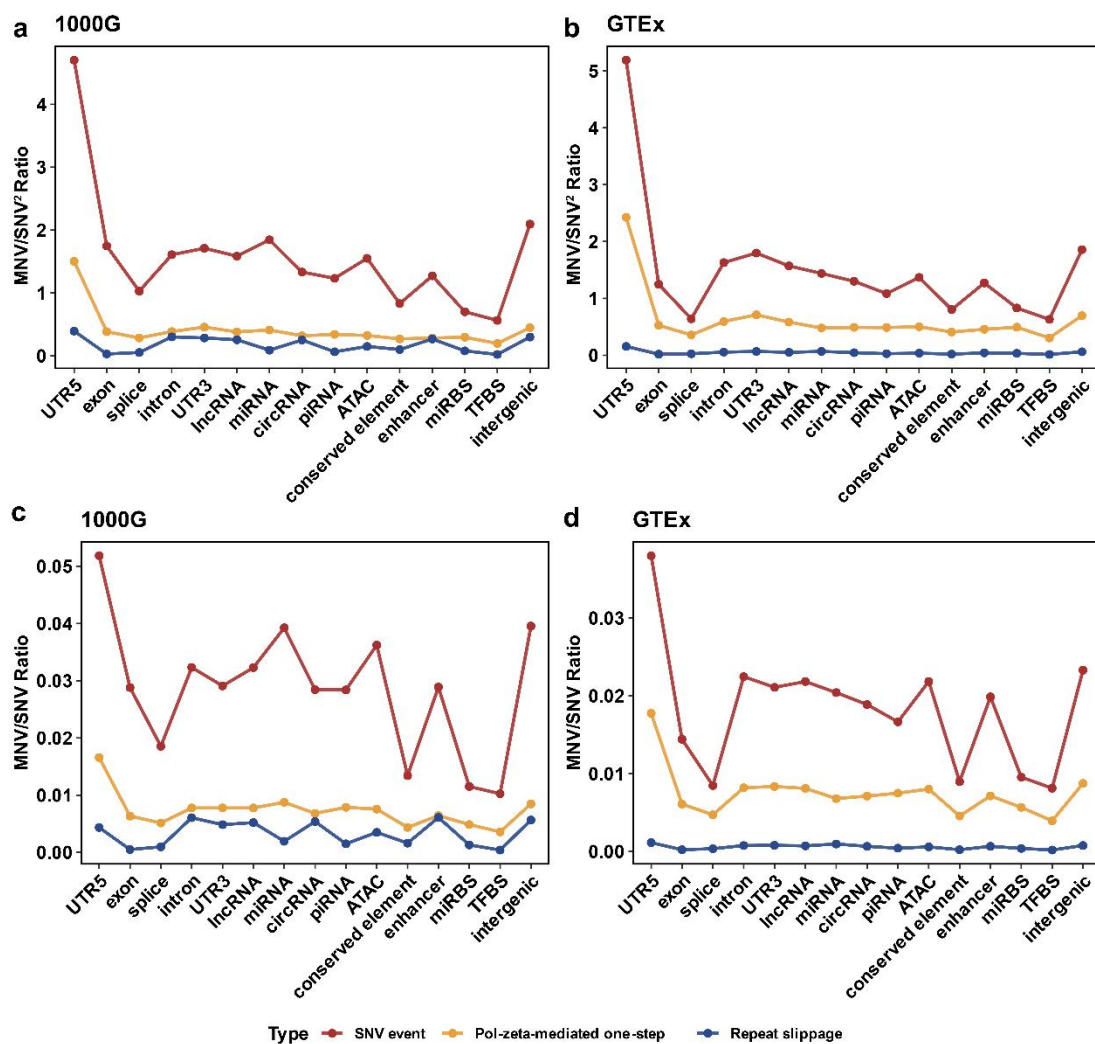


Figure R1. Comparison of MNV density using different normalization methods across functional genomic categories. (a-b) MNV density normalized by the square of the SNV rate (MNV/SNV²) for the 1000G (a) and GTEx (b) datasets; (c-d) MNV density normalized by the SNV rate (MNV/SNV) for the 1000G (c) and GTEx (d) datasets.

In the revised manuscript, we have modified the sentences in lines 201-204 as

follows:

“Highly conserved regions (exon, splice region, conserved element, miRBS, and TFBS) exhibited significantly lower MNV/SNV ratios, which could reflect an evolutionary constraint against MNVs in these functionally critical regions. regions 22,”

2) Check whether difference with gnomAD MNVs could be explained by inclusion of low complexity region

I doubt that “This discrepancy arises primarily because our study incorporates the 1000G project (with samples from five populations), whereas gnomAD is dominated by European populations.”, as the gnomAD v2 dataset also includes a good number of non-european samples (e.g. it contains >9,000 East Asian samples, far more than that in 1000G <https://www.nature.com/articles/s41586-020-2308-7/figures/1>).

Could this be rather due to differences in the variant filtering? In particular, considering that in gnomAD, MNVs in low complexity regions were conservatively filtered out (as stated in the manuscript <https://www.nature.com/articles/s41467-019-12438-5>: “These findings come with the caveat that variants in repetitive regions will have higher error rates due to slippage and misalignment errors, but we have reduced this risk by applying random forest filtering for individual sites, as well as removing all the variants in low-complexity regions from our analysis”), I wonder a large majority of those MNVs uniquely discovered in this study are actually in low complexity regions and are filtered out in gnomAD (or just generally relatively low-quality variants that are filtered out in gnomAD).

Could the authors check whether this is the case? If it is, that is still fine as long as that is clearly mentioned. Other than that, I fully agree with the added value of 3-5 joint MNVs.

Response: Thank you very much for correcting our previous misjudgment on the ethnic composition of the gnomAD dataset and providing key reference literature on gnomAD's variant filtering strategy. Therefore, we re-analyzed the newly identified 2-joint MNVs from the 1000 Genomes dataset and calculated their proportion in low-complexity regions (LCRs, identified via the symmetric DUST algorithm). The results showed that approximately 68.37% of these novel 2-joint MNVs are located in LCRs, likely representing the primary contributor to novel MNVs. As a major goal of this work was to construct a comprehensive MNV resource, we did not apply overly stringent filtering criteria for MNVs.

Other factors underlying the discrepancy in MNV counts include differences in sample compositions between this study and gnomAD (each sample could contribute some novel MNVs) and our inclusion of 3- to 5-joint MNVs in the analysis.

In the revised manuscript, we have added the following sentences in Supplementary materials in lines 6-7:

“Filter phased VCF: considering the sensitivity or accuracy of phasing and the range of regulatory elements (e.g., miRBS and TFBS), MNVAnno will filter out variants that exist alone within a certain window (default 10 bp). In this study, to construct a relatively comprehensive resource of MNVs, we did not filter MNVs located in low-complexity regions (LCRs).”

3) Provide additional interpretation of deviation of delta MFE from additive

Figure R10 is beautiful. Is there a way to quantify whether these deviations are more than expected by chance? If that is hard (maybe it is, since we cannot get the null distribution), I think mentioning the direction of the non-additive effect in Figure R10 is meaningful here, to show that it is not random and to interpret it biologically. Specifically, I see the mitigation effect (i.e. when sum of SNV effects are large, MNV effects tend to be smaller than the sum, and vice versa), so in many cases, we expect

(effect of one SNV) < (effect of MNV) < (sum of the effects of constituent SNVs), if I am interpreting the figure correctly. Mentioning this observation together with an interpretation would be valuable.

Response: Thank you very much for the positive feedback on Figure R10 and the insightful hypothesis about the non-additive effects of MNVs (mitigation effect). To further investigate the relationship between MNV effects and the sum of constituent SNV effects, we plotted a scatter plot comparing MNV effects with the sum of constituent SNV effects (Figure R2). The slope of the regression line was 0.85 (standard error = 0.015; 95% confidence interval: 0.82–0.88), which is significantly less than 1 (one-sided *t*-test, $p = 7.4 \times 10^{-22}$). When the absolute value of the sum of SNV effects is large, the predicted effect of MNV tends to be smaller than the sum, suggesting a mitigation effect of MNVs.

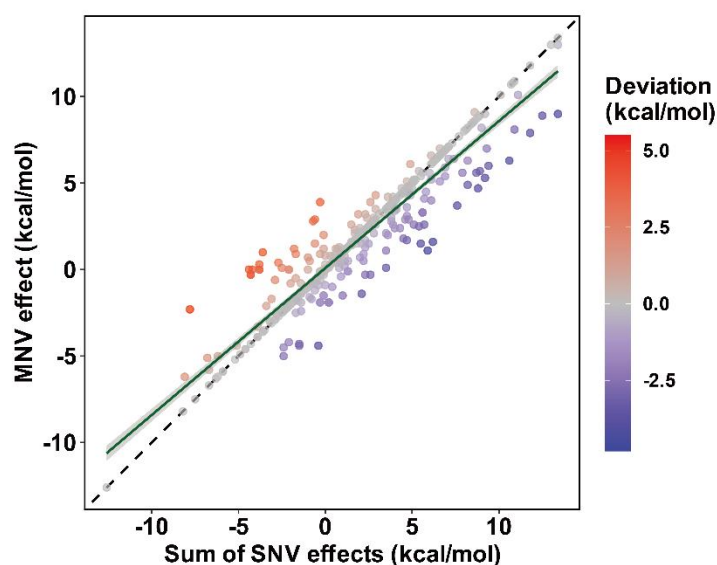


Figure R2. Comparison of MNV effects versus the additive sum of constituent SNV effects on RNA secondary structure stability. The black dashed line ($y = x$) denotes the perfect additive model, while the solid green line with grey shading indicates the linear regression fit and 95% confidence interval. Points are colored by deviation magnitude.

We further quantified the proportion of MNVs exhibiting the specific pattern: (effect

of a single constituent SNV) < (effect of MNV) < (sum of constituent SNV effects). Specifically, among MNVs with effects deviating from the sum of their constituent SNV effects, 25.81% exhibited this pattern. Moreover, when the absolute value of the sum of constituent SNV effects exceeded 5 kcal/mol, 78.13% of such variants displayed the same pattern. Given that the energy changes were derived from computational predictions, additional experimental validation is required to fully substantiate this observation. For this reason, we did not include this specific pattern in the main text.

In the revised manuscript, we have added the following sentences in lines 272-277.

“The results revealed that 52.16% of MNVs exhibited Δ MFE values deviating from the sum of their constituent SNVs' Δ MFE (Supplementary Fig. 12a), indicating non-additive effects on ncRNA secondary structure stability. Furthermore, we found that when the absolute value of the sum of SNV effects is large, the predicted effect of MNV tends to be smaller than the sum, suggesting a mitigation effect of MNVs (Supplementary Fig. 12b)”

Reviewer #2:

The authors have responded to my comments and implemented meaningful improvements to the article. The results about the non-additivity are particularly interesting.

One minor issue is the inconsistency in the number of significant MNVs reported (232 in the response vs. 233 in the revised manuscript), which should be clarified and corrected.

Once this is clarified, I support acceptance for publication.

Response: Thank you very much for your positive assessment and for supporting the acceptance of our manuscript. We also appreciate your keen attention to detail in

spotting this numerical inconsistency.

We have verified the count of significant MNVs (confirming the correct value as 233). The number "232" in our previous response letter was a typographical error. We sincerely apologize for this oversight and have thoroughly checked the entire manuscript and supplementary materials to avoid similar inaccuracies.

Response to the reviewers' comments

Reviewer #1 (Remarks to the Author):

I truly appreciate the authors' additional efforts. My questions are well addressed. I have one final comment regarding my previous question "2) Check whether difference with gnomAD MNVs could be explained by inclusion of low complexity region", as detailed below.

If the authors agree on the interpretation, adding/editing a few sentences in the main text before final publication would be appreciated.

I do not have other additional concerns/comments. Thank you so much again for working on all the revisions and congratulations on the completion of this amazing work!

Regarding "2) Check whether difference with gnomAD MNVs could be explained by inclusion of low complexity region":

Thanks a lot for performing additional analysis to identify that the inclusion/exclusion of low complexity regions (LCRs) explains a significant amount of the difference in the number of MNVs identified in gnomAD vs this study. While I agree that including LCRs to construct a more comprehensive list of MNVs is a rational choice, I think the fact should be more explicitly stated in the main text and methods, not only in the supplementary text.

I appreciate it if the authors could add a few sentences clarifying the facts that

- Approximately 70% of the "novel" joint MNVs are in LCRs
- LCRs are known to be enriched for sequencing errors, so the "novel" joint MNVs could contain some fraction of false-positives
- The users should take caution when interpreting these joint MNVs in LCRs
- (but the authors included them nevertheless, to make the list more comprehensive)

Although I will let the authors decide where and how to include these statements, it

seems that the places such statement could be included are around:

- Lines 134-135 "...2,222,090 MNVs were newly discovered compared to those in gnomAD..."
- Lines 365-366 "...found 2,222,090 novel ones compared with previous Wang's study utilizing gnomAD dataset..."
- Lines 410 "Identification and annotation of MNVs" section of the method
- Figure 2a legend

Response: Thank you very much for your positive feedback on our previous revision. We sincerely appreciate the suggestions you provided in this third round of review, and we have made the corresponding revisions accordingly.

1) In the Results section (Lines 133-135), we have revised the description as follows: "Among them, 2,222,090 MNVs were newly discovered compared to those in gnomAD¹³ (Fig. 2a), and approximately 70% of them are in low-complexity regions (LCRs). "

We have included this clarification in the main text (immediately adjacent to the Fig. 2a citation tag). Therefore, we did not add additional statements in the figure 2a legend, as this would lead to redundancy and inconsistency with the surrounding text.

2) In the Discussion section, we have modified the sentences in Lines 365-366 and 389-394 as follows:

"In this study, leveraging genotype data from 700,134 samples, we identified 3,984,258 MNVs and expanded the human MNV list to 8,199,654."

"However, in our study, to build a comprehensive human MNV list, we retained MNVs in LCRs, a key filtering parameter difference from gnomAD-based analysis⁹ and one of the main sources for our identification of novel MNVs. LCRs are known to be enriched for sequencing errors, which may lead to a fraction of false positives; therefore, the verification of MNV existence and the elucidation of the precise

functional mechanisms of MNVs still require further experimental validation.”

3) In the Methods section, we have modified the sentences in Lines 416-418 as follows:

“We first selected variants that appeared in the same individual, in cis, and within 10 bp distance in phased VCF files by MNVAnno (See Supplementary Materials for details). To generate a relatively comprehensive MNV resource, we did not filter out MNVs located in LCRs.”

I do agree that other factors such as sample composition, inclusion of 3- to 5- joint MNVs are also contributing factors. Just want to make sure that the readers understand that different filtering criteria is one of the main sources of the difference in the number of MNVs discovered, and won't get over-excited about the fact that we are discovering so many additional MNVs. Thanks a lot.

Response: We sincerely appreciate your recognition and valuable suggestions. As you recommended, we have added relevant content in the manuscript and explicitly highlighted this point in the Discussion section — specifically clarifying that different filtering criteria are one of the main sources of discrepancies in the number of MNVs discovered, to ensure readers will not overstate the significance of the additional MNVs we identified. Thank you again for your time, careful review, and insightful comments.