Dear Diego A. Hartasánchez,

First, we would like to thank you and both reviewers for their constructive comments about our manuscript together with their positive feedback about our choice to integrate and discuss the entire dataset generated from this study even if some results were not as conclusive as expected. We believe that this work well reflects the challenge of deciphering the genetic bases of insecticide resistance in insect vectors and well fits with the inclusive open-minded PCI editorial policy, which we strongly support.

You will find below our responses to reviewers' comments together with changes made to the revised version of the manuscript. The revised manuscript and supplementary files have been deposited on bioRxiv under BIORXIV/2024/587871 – V3 version (https://doi.org/10.1101/2024.04.03.587871b)

The pdf of this 'V3 revised version' together with all associated supplementary files can be downloaded at https://www.biorxiv.org/content/10.1101/2024.04.03.587871v3.

All line numbers mentioned in our answers to reviewers refer to the .docx 'track change revised version' submitted to PCI Evol Biol.

We sincerely hope that the significant changes made to the manuscript will make this study suitable for its recommendation by PCI Evolutionary Biology and that this study will be of interest for the scientific community.

Best wishes,

Jean-Philippe DAVID, on behalf of all co-authors.

Decision for round #1 : *Revision needed*

Major revision suggested

Dear Tiphaine Bacot and co-authors,

First of all, thanks for submitting your work to PCI Evolutionary Biology, and please forgive the delay in getting back to you with the reviews for your manuscript.

I enjoyed reading your manuscript and find that it's a research of great importance. I have received two reviews by experts in the field, which are also optimistic about your work. The reviewer's comments are mostly positive with respect to the importance of your work, and the quality of your manuscript, acknowledging the complications in performing this type of

research in a challenging genome. They do, however, raise important points that would need to be addressed prior to my recommendation.

Please, address all of the reviewers' points. In particular, the findings of Cattel et al., 2020, should be encompassed adequately in the manuscript; apparent inconsistencies between figure 1 and table 1, and between figures 1 and 4, need to be clarified; and I do suggest reevaluating the substructure of the results section to see if you find a way to present the results in a way that helps clarify some of the unclear points of the manuscript.

Finally, although I consider that all of the reviewers comments and questions should be addressable, please feel free to explain your point of view if you happen not to agree with them.

I'm looking forward to your revisions.

Best regards,

Diego A. Hartasánchez

Review by Diego Ayala (Reviewer 1)

The manuscript by Bacot et al., presents a thorough and well-articulated study on the characterization of a genomic duplication that implement insecticide resistance in Aedes aegypti, a major vector of arbovirus across the World. The research is meticulously conducted, with a large number of different experiments and testing-hypothesis, with a clear exposition of methodologies and results. The authors look at to demonstrate the functional implications of the genomic duplication, supported by robust experimental data. The discussion provides insightful interpretations and situates the findings within the broader context of insecticide resistance mechanisms in this mosquito. It is clear that other similar genomic characterization of this duplication would help to udnerstsand its origin and spread. Overall, the clarity and coherence of the narrative, combined with the significance of the research, make this article a valuable contribution to the field and worthy of publication.

R1: My unique concern is about how the authors integrate the previous results (Cattel et al., 2020) in the present study. The duplication was already observed and roughly characterized. However, even if they mention this fact in the introduction (i.e. line 97), they perform some analysis as if they knew little about the duplication. For instance, I wondered why they performed comparative analysis on F1 and F2 (lines 144 and so on) to understand if there is any maternal effect. It is now obvious when you see later that this duplication is in the chromosome 1. Moreover, they already knew that the duplication affected multiple P450s, however it seems as they discovered that in the paper. In my opinion, if the authors provide former information about their previous results, the paper will be easier to follow up.

> We thank reviewer 1 for this interesting comment that calls for a better explanation of the rationale of the study in regards of our previous work published as Cattel et al 2020. First, we would like to emphasize that although this previous CNV screening evidenced the potential

presence of a duplication at this P450 locus, the multi-resistant nature of the Guy-R composite population used, carrying multiple resistance alleles to different insecticides (including the three major Kdr mutations Val410Leu, Val1016Ile or Phe1534Cys at high frequency and several metabolic alleles) did not allow an in depth study of the association of these P450s with deltamethrin resistance. In the present study, the more specific deltamethrin resistance phenotype of the isofemale resistant lines used together with their deprivation from the three major Kdr mutations mentioned above represented an opportunity to better characterize the role of this P450 duplication in deltamethrin resistance. In addition, the discontinuous nature of the sequencing data used in Cattel et al 2020 (targeted exon pool-sequencing) together with the high complexity of Ae. aegypti genome (notably its high density in repeated regions and transposable elements) did not allow resolving the genomic architecture of the duplicated locus at that time. The introduction of the manuscript (lines 95 to 106) has been thoroughly modified to better reflect these points and better introduce the study in regards of previous work.

In regards of the mode of inheritance of the P450 duplication, its co-localization with the sex locus on the chromosome 1 in a position neighbouring a large chromosomal region showing a low recombination rate and a high genetic differentiation between males and females (Fontaine et al., 2017) lead some interrogations about its mode of transmission. As such question was not investigated in Cattel et al 2020, we took benefit of the present study to conduct reciprocal crosses to confirm the absence of sex-linkage or maternal effect affecting this resistance allele. Though of minor importance, this section is now better introduced early on in the manuscript (lines 158-159) and then shortly recalled in the discussion section.

R1: A second, but minor question, is about the role of the duplication contributing to insecticide resistance. The article provides a comprehensive analysis of genomic duplication conferring insecticide resistance but occasionally presents controversial results regarding the duplication's role (i.e. lines 347-348, line 401). Clarifying the transitions between the results would enhance the manuscript's overall clarity.

> Indeed, this is the delicate point of the manuscript as the P450 duplication is significantly associated with deltamethrin survival in our F2 association study and previous studies (including Cattel et al 2020 and others), while RNAi and experimental evolution data are less conclusive, likely because of technical issues and the presence of the Kdr 1011Met resistance allele (see discussion about resistance costs). The manuscript has been modified to better reflect this as follows: Lines 401-402: this sentence has been modified in order to not focus on a particular P450 gene but rather to state about the role of the whole duplicated locus in the P450-mediated resistance phenotype observed. Line 445-446: this sentence has been modified in order to better distinguish the association of the P450 duplication with deltamethrin survival observed in F2 individuals from its unexpected response to artificial deltamethrin selection in presence of the Kdr 1011Met resistance allele.

R1: Overall, the paper is very complete and of a high scientific level, with numerous assays spanning from phenotype to genotype. Congratulations to the authors for their meticulous work and significant contributions to the field of insecticide resistance research in general, and in Aedes aegypti in particular.

> We thank reviewer1 for this positive comment about our work and hope that our answers and the significant changes made to the revised version of the manuscript will be satisfactory.

Review by anonymous (reviewer 2)

This study by Bacot, Haberkorn *et al.* focuses on a genomic duplication encompassing multiple P450 genes suspected to be associated with insecticide resistance in French Guinana *Aedes aegypti* mosquitoes. The authors used a wide array of approaches (bio-assays, molecular quantification through ddPCR, RNA interference, experimental evolution and both long and short-read sequencing) to assess the resistance levels associated with this duplication, its fitness cost, and uncover its genomic architecture. This study is thorough in its investigation and highlights the challenges faced when identifying new copy number variants associated with xenobiotic resistance, most of all in a genomic environment as challenging as the *A. aegypti* genome. Overall, this article is well written and concise despite the breadth of analyses performed, but in our opinion a few key points need to be addressed carefully.
The results would greatly benefit from few clarifications, the framework imposed by the impossibility to study the effect of the massive duplication independently from the 1011Met allele really complexifies the understanding of the analyses performed here. Because of various inconsistencies across the analyses we have the feeling that we were not able to fully assess the quality of the study.

R2: For instance some F1 crosses are reported in Figure 4 quite late in the manuscript but not in the first paragraph that refers to supplementary file 1. Why couldn't the authors use the results of the first analyses?

> The co-localization of the P450 duplication with the sex locus on the chromosome 1 in a position neighbouring a large chromosomal region showing a low recombination rate and a high genetic differentiation between males and females (Fontaine et al., 2017) led some interrogations about its mode of transmission. This rationale is now better introduced in the result section (lines 158-159) and further mentioned in the discussion.

In this frame, the experiment presented early on in the manuscript in Supplementary Figure 1 only aimed at deciphering the mode of transmission of resistance alleles present in the resistant IR0F line by performing both 'Bora-Bora x IR0F' reciprocal crosses and comparing the resistance level of F1/ F2 offspring obtained from each cross. The deltamethrin bioassay used for this thus required an insecticide doses providing a good dynamic range to detect mortality variations between individuals (of each sex) generated from each reciprocal cross. As males Ae. aegypti naturally show a lower tolerance to deltamethrin, different insecticide doses were chosen for males and females in order to keep mortality in a range of 30-50% in F1/F2 offspring for each sex. Such mortality level provided a good dynamic range to detect mortality variations between F1 and F2 offspring from each cross as opposed to a higher dose of insecticide that would have led to higher mortality values in F1/F2 offspring and saturation effect. This point is now better explained in the method section (Lines 530 to 545).

On the other hand, the association study on F2 individuals presented in Figure 4 was performed as a separate experiment. As no sex/maternal transmission bias was previously detected, this experiment was therefore performed on females only obtained from both reciprocal crosses (equal number of females from each cross) and with a higher dose of insecticide (0.03% for 1h as mentioned in the method section) in order to segregate most resistant individuals (>80% mortality).

R2: The authors define the resistance as semi-dominant while the mortality in F1 crosses in Figure 4 is above 80%. More surprisingly the mortality in the F2 crosses is also close to 80% when one could have expected it to decrease because of the presence of resistant homozygous genotypes.

> As stated above, the association study presented in Figure 4 was performed using a higher dose of insecticide in order to segregate most resistant individuals (>80% mortality). The semi-dominant behaviour deduced from data presented in Supplementary Figure 1 only implies that heterozygous will show and intermediate resistant phenotype as compared to homozygotes, which is clearly the case and expected in the case of a gene duplication associated to a quantitative gene expression variation. Such intermediate phenotype of F1/F2 individuals is not so evident in Figure 4 due to mortality saturation in the susceptible line 'Bora-Bora'.

> The frequency of the P450 duplication is not expected to change from F1 to F2 though the F1 progeny is likely mostly composed of heterozygous and the F2 progeny is likely composed of a mix of the three genotypes (expected proportion ¼ ; ½ ; ¼). As mortality data are inferred from pool of individuals it is not be possible to infer the putative impact of the presence of homozygous duplicated individuals on the overall mortality between F1 and F2, especially in the presence of the Kdr 1011Met mutation. In this way, we acknowledge that it would be of interest to assess the individual genotypes of F2 dead and survivors for both resistance alleles, though this was not possible given the nature of the samples used for digital droplet PCR quantification (pools of individuals).

R2: Another major concern comes from the fact that the writing implicitly suggests that the massive duplication encompassing the CYP genes provide resistance while, and as nicely acknowledged by the authors, they obtained contrasted results regarding this point. For instance, the analyses presented in figure 1 are pretty convincing about it (but see major comment below) but there is no further support for it in the other analyses.
In figure 4, the slight difference in copy number between F2D (dead, 4) and F2S (survival, 3.3) let's suppose that the dead ones were globally heterozygous for the duplication (mean number of copies 3.5 in the F1 crosses). How could the authors then rule out that the survivors are simply mostly homozygous for the Met too?

> As a point of detail, Figure 4 shows that the mean CYP6BB2 copy numbers estimated by ddPCR are 3.35 and 4.0 for F2D (dead) and F2S (survivors) respectively with CI intervals based on the number of positive nanodroplets and the Poisson statistic distribution used for ddPCR quantification. Indeed, these ddPCR data obtained from pools of mosquitoes do not allow to infer the individual genotype of dead/survivors though such individual data would be of interest to better understand how resistance alleles eventually co-segregate in dead and

survivors. Therefore, this result section does not state that F2 deads are mostly heterozygotes for the P450 duplication, nor survivors being mostly homozygotes for the kdr mutations… This section only states that the frequency of both alleles (P450 duplication and Kdr 1011Met) is higher in survivors as compared to deads, with F1 individuals (likely mostly heterozygotes for both alleles) being in between. Considering that IR0F individuals shown in Fig4 are mostly homozygote for the duplication and the Kdr 1011Met allele (allele frequency of the heterologous duplicated allele = 0.5), then figure 4 suggests that not all F2 survivors are homozygous for each allele as their mean CNV and Kdr 1011Met frequency are both below those observed for the IR0F line.

R2: Here again, the frequency of the Met increases between the F2D and the F2S. In absence of individual genotyping showing that the Met and the duplication segregated independently (despite the fact that they are on different chromosomes) it is hard to discriminate the respective effects.

> As stated above, the nature of the ddPCR data provided does not allow inferring the individual genotypes of dead/survivors in this experiment though such individual data would be of interest to better understand how both resistance allele eventually co-segregate in dead and survivors. However, as both resistant alleles are located on distinct chromosomes we believe that it is reasonable to think that they segregate randomly from F1 to F2 offspring. In that line, the experimental evolution data presented in Figure 5 clearly do not support any genetic linkage between the two resistance loci.

R2: The order of the analyses and the choice of sub-heading does not help either and we suggest that the authors revise the structure of the results; we list below the sub headings in order of appearance in main text:

- Deltamethrin resistance is associated with P450 activity
- The resistance phenotype is autosomal and semi-dominant
- The resistance phenotype is associated with a duplication affecting multiple P450s
- Genomic architecture of the duplication
- The P450 duplication is associated with deltamethrin survival
- Multiple P450s carried by the duplication may contribute to deltamethrin survival
- The duplication is hardly retained by selection in presence of the 1011Met kdr allele

> We tried to improve the structure of the manuscript and rewording some section heads. After turning things in different ways…, we finally decided to adopt the following structure which in our view improve the clarity of the manuscript while keeping things in a logical order.

Results:
- Deltamethrin resistance is associated with P450 activity and is autosomal
- Overexpression of the duplicated CYP6 gene cluster
- Genomic architecture of the duplicated loci
- The P450 duplication and the Kdr 1011 mutation are both associated with deltamethrin survival
- Multiple CYP genes carried by the P450 duplication may contribute to resistance
- The P450 duplication is hardly retained by selection in presence of the Kdr 1011 allele

Discussion:
- The P450 duplication affects a cluster of genes previously associated with resistance
- Contrasted architectures of the duplications affecting the P450 and the Kdr loci
- The P450 duplication has a limited adaptive value in presence of the Kdr 1011 mutation
- Conclusions

R2: Additional comment:
- Providing a figure early in the main text presenting the different lines and crossing design used in the different analyses would make the results easier to follow.

> In order to clarify this point for the reader, two sub-figures showing the crossing designs associated with the F2 genotype-association study and the experimental evolution study have been added to Figure 4 and Figure 5. The reciprocal crossing design associated with the study of the mode of transmission of resistance is also shown on Supplementary Figure 1.

R2: Figure 1 is essential in showing the implication of P450s duplication in resistance. It is the only clear support for it through the whole study. However, there seems to be a strong discrepancy between the IR13 resistance status indicated in table 1 and the results shown in fig. 1. IR13 is described as providing low resistance (ca. 95% mortality rate) in the table but is later on referred to as susceptible in the main text. Surprisingly, the same line, IR13,displays a ~45% mortality rate without PBO in fig. 1, despite a lower exposure to the insecticide. Even if it is not statistically significant, IR13 shows a reduced resistance when exposed to PBO too

> We understand that the different mortality levels shown for the IR13 line and the terms used to define its susceptibility level to deltamethrin in Table 1 and the rest of the manuscript were a bit confusing for the reader. Indeed, the IR13 line shows a high susceptibility to deltamethrin as shown in table 1 using 0.05% deltamethrin for 40 min exposure (as mentioned in Table 1 caption). This high susceptibility of the IR13 line to deltamethrin was also confirmed by Epelboin et al. in 2021 using a lower diagnostic dose of 0.037% for 1h where the IR13 line was found susceptible to deltamethrin. The IR13 line still shows a lower susceptibility to deltamethrin than the fully susceptible laboratory strain Bora-Bora (which has been maintained for decades in insectaries) as frequently observed with field-derived susceptible populations as compared to established laboratory lines.

As stated in Figure 1 caption (and in the methods section), the dose of deltamethrin used for PBO-bioassays presented in Figure 1 was lower (0.03% for only 30 minutes) in order to keep mortality level in PBO-exposed individuals in the dynamic range for both IR13 and IR0F lines (avoid mortality saturation effect following PBO exposure that would impair mortality comparison between the two lines). These different doses explain the apparent discrepancies noted by reviewer2 between Table 1 and Figure 1. In regard of the low effect of PBO exposure on the IR13 line, such lower (non-significant) effect is not surprising as even susceptible lines show a background P450 activity (associated to a basal insecticide degradation rate). What is important to note here is the stronger increased mortality observed after PBO exposure in the IR0F line as compared to the IR13 line which clearly supports an elevated P450-mediated detoxifying activity in the resistant line.

Considering this and in order to clarify the status of the IR13 line for the reader, the IR13 line is now described as 'susceptible' in Table 1 and in the whole manuscript. A short sentence better describing the susceptibility status of the IR13 line as compared to the laboratory

strain Bora-Bora has also been added early on in the first result section. In addition, the justification of the different insecticide doses and exposure times used for each experiment is now more apparent in the method section dedicated to bioassays.

R2: Could the authors specify early in the manuscript the genotype of the various strains/lines regarding the presence of the duplication too? Could the authors consider moving Fig. 1 and the corresponding section after figure 2?

> As the IR lines used in the present manuscript are isofemale lines (i.e. obtained from a single female that may have been fertilized by multiple males) inferring early on their 'genotype' for resistance alleles would be somewhat misleading. In addition, we would prefer to not move figure 1 (dealing with the overall resistance phenotype) after Figure 2 which goes deeper into the biochemical/molecular resistance mechanisms.

R2: In fig. 2, the presence of the fifth overtranscribed P450 gene is scantly discussed. Was the position of this gene investigated after the *de novo* assembly? Would it be possible that this gene is mis-positionned and is in fact a part of the duplication encompassing the four other CYP genes? Knowing that this gene show both a higher transcription level and an increase in gene dosage, how could the author exclude its effect on the resistance phenotype.

> This gene (AAEL028635) was not initially annotated as a P450 in the Ae. aegypti genome at the time of the study but a BlastP search against NCBI Refseq clearly identified it as a putative CYP9. This gene is part of a large CYP9 cluster located on chromosome 3 which are phylogenetically divergent from the CYP6 genes identified on Chromosome 1. Therefore, this gene is not likely belonging to the CYP6 duplicated cluster under study.

Although this CYP9 gene shows a significant over-transcription in resistant lines, its transcription profile does not fully match with a deltamethrin resistance related-gene with transcription ratios being higher between resistant lines versus the susceptible IR13 susceptible line (~4 fold) than between resistant lines versus the fully susceptible Bora-Bora line (~2 fold) suggesting a differential transcription levels in susceptible lines. In terms of CNV, this gene also shows an inconsistent CNV across the four comparisons with only 1.2 fold and 1.07 fold in the IR03/Bora-Bora and IR03/IR13 comparisons respectively, leading us to not retain it as a CNV associated with deltamethrin resistance in the present study (see our experimental design integrating the four distinct line comparisons in the method section). This point is now better described in the result section (Lines 202 to 205).

Overall, though it cannot be excluded that this CYP9 gene is associated with insecticide resistance, its expression and CNV profiles show an unclear association pattern in our deltamethrin resistant lines. In addition, its sequence and clear positioning on chromosome 3 does not support it as being a member of the CYP6 duplication on chromosome 1. Having said this, it cannot be excluded that in addition to the kdr mutation 1011Met, the over-expression of this CYP9 also respond to deltamethrin selection in our experiment evolution experiment. This point in now acknowledged in the discussion (lines 472-473).

R2: The authors report that the two resistant lines show a ca. 2 - fold increase in DOC for the duplicated region but these data are not shown. The heatmap let room for

interpretation for the actual change in relative DOC. Would it be possible that the IR13 line is in fact heterozygous for the duplication ? Could the authors provide a table with the actual normalised relative depth for each of the CYP trapped into the duplication to the depth of the corresponding CYP for both "susceptible strains".

> The presence of the CYP6 duplication in the IR13 line is neither supported by CNV quantitative data nor by the genomic architecture of the P450 duplicated locus deduced from (short and long) sequencing reads. In support of this, supplementary Table 2 provides detailed CNV data for all genes identified by the present study including 'raw exonic read count', 'normalized exonic read counts' and 'CNV fold changes' for the four resistant Vs susceptible comparisons. Finally, long read sequencing of the IR13 line did not identify the presence of the PiggyBac-like transposon and the Piggy/HAt chimeric region associated with the P450 duplication.

Minor commentaries:

R2: Is it common (or even meaningful) to list second co-equal contributions, or is it a typo and these authors are all first -co-equal contributions? We have the feeling that such a level of details in the authorship does not match with the publishing model put forward by PCI, but we'll let the recommender decide upon that.

> The results presented here were obtained through >5 years of team work (from 2018 to 2024) and involved a lot of staff and Master internships. Given this, it was not ethically possible to distinguish the relative input of some authors which led us to all assign them as co-second author (Haberkorn, Guilliet, Cattel, Kefi). We don't think this goes against PCI editorial policy as this clearly denotes a collaborative research effort.

R2: L154 IR13 is either called susceptible or slightly resistant (see Tab. 1) please harmonise.

> In order to improve overall manuscript clarity, the IR13 line is now mentioned as susceptible in the whole manuscript including Table 1 (see justification above).

R2: L154 "Among the 11078 genes detected", what is meant by "detected"? Does it mean DE genes in total or total number of genes with enough transcript sequenced in all the lines? Also please provide the total number of genes in *A. aegypti* to give a reference frame.

> A total of 13614 protein coding gene were studied as stated in the method section. Among them total of 11268 (82%) protein-coding genes (not 11078 gene as previously shown) were detected in adult females by our RNA-seq analysis (i.e. passing the minimum coverage filter described in the method section). This mistake has been corrected in the manuscript and the percentage added (lines 171 and 594).

R2: L 195 "Indeed, both short read and long read data revealed that this Ile1011Met mutation …" This is the first instance where short reads are mentioned, the results would gain in clarity if a little bit more of methods was added here, or consider moving the methods before the result section.

> This has been clarified in the result section (lines 194 to introduce whole genome short read data). Long read sequencing are introduced afterward in the section dealing with the genomic architecture of the two duplicated loci.

R2: L 232 "(Fisher test p value = 0.035)". This is surprising, please add the value of the statistics and the degrees of freedom for all the tests performed.

> Here is the way we performed this simple test (now reported on line 257)

```
mat
#               dupli   out
# biallelic     1336    1334
# polyallelic   5       0

fisher.test(mat, alternative = 'less')
        # Fisher's Exact Test for Count Data
        #
        # data:  mat
        # p-value = 0.03154
        # alternative hypothesis: true odds ratio is less than 1
        # 95 percent confidence interval:
        #  0.000000 0.823533
        # sample estimates:
        # odds ratio
        #          0
```

R2: L 238 replace "segregated with a high dose" with "exposed to a high dose"

> the term 'segregated' was replaced by 'exposed to' in this section (line 265, 272, …).

R2: L. 244 and 249: please provide statistics on the copy number differences and Met allele frequencies between survivors and dead individuals.

> These data have been generated by digital droplet PCR where each reaction is composed of >15000 droplets. With these very high numbers, and based on the positive/negative droplet ratios, the Poisson statistic Law allows i) a correction of raw values to estimate the initial target concentration and ii) the calculation of confidence intervals (see https://www.bio-rad.com/webroot/web/pdf/lsr/literature/Bulletin_6407.pdf). We used the default 95% confidence level. It is clear on Fig 4 that confidence intervals do not overlap between dead and surviving F2 individuals for both CYP6 copy number and Met allele frequencies, which is a sufficient (but not even necessary) condition to state that the values are different at this confidence level.

R2: L. 254 please define N≥5

> This implies that at least 5 replicates WHO test tubes of 20-25 adult females were used for each condition in this bioassay, as described in the method section.

R2: Figure 4: The distribution of error bars is surprising given the difference in sample size (113 vs 15), please check.

> As mentioned above, these molecular data were generated from pools of mosquitoes using digital droplet PCR, the error bars (95% CI) were estimated using the Poisson Law based on

the number of positive/negative droplets and do not take into account the number of mosquitoes within each pool.
With a low number of mosquitoes (15 F2D), a sampling bias cannot be excluded. A strong sampling bias would jeopardize the conclusion [that copy number is higher in F2S than F2D], but this is unlikely in our case a supported by this 'coin de table' simulation:

*Let's assume N=15 (number of F2S individuals) with the following genotype proportions (2:1:0 for DD:Dd:dd, in other word p=0.33) and a number of copies (4:3:2 for DD:Dd:dd). The real mean copy number (4 for DD and 3 for Dd) = 4*p+3*(1-p) =3.33. Now assume that instead of giving us 5 heterozygotes (N*p), the devil gave us more.*

sampled_heterozygotes	proba	deduced_CN	CN_error
5	2.143071e-01	3.333333	0.00000000
6	1.785892e-01	3.400000	-0.06666667
7	1.148074e-01	3.466667	-0.13333333
8	5.740368e-02	3.533333	-0.20000000
9	2.232365e-02	3.600000	-0.26666667
10	6.697095e-03	3.666667	-0.33333333
11	1.522067e-03	3.733333	-0.40000000
12	2.536779e-04	3.800000	-0.46666667
13	2.927052e-05	3.866667	-0.53333333
14	2.090752e-06	3.933333	-0.60000000
15	6.969172e-08	4.000000	-0.66666667

*(sampled_heterozygotes=i, proba= dbinom(i, N, p), pi=i/N, deduced_CopyNumber= 4*pi+ 3*(1-pi), CopyNumber_error= real_CN – deduced_CN)*

Then the simulation shows that it takes an excess of 5 heterozygotes (i=10) to reach a ~10% overestimation of F2S average copy number. Our claim would still hold, and the probability of a such excess is already very low (<1%). The cumulated probabilities of an even higher excess is below 0.2% and, with a maximal overestimation of 20%.

R2: It would be nice to have direct access to the SRA repository through the use of Hyperlink

> The Data availability section provides accession numbers to all sequencing data used in the present study.