1. *Brief Overview of the study*

a) Title of the paper and full reference

"The Impact of Transposable Elements on Tomato Diversity"

Domínguez, Marisol, Elise Dugas, Médine Benchouaia, Basile Leduque, José M. Jiménez-Gómez, Vincent Colot, and Leandro Quadrana. 2020. "The Impact of Transposable Elements on Tomato Diversity." *Nature Communications* 11 (1). https://doi.org/10.1038/s41467-020-17874-2.

b) Summary

a. Background and aims

Cultivated tomatoes have extensive phenotypic variation despite having low genetic diversity due to a bottleneck associated with the migration of the tomato from Peru to Ecuador and/or from Mesoameria to Europe. The high phenotypic variation in cultivated tomatoes is therefore thought to be due to large effects from a few rare alleles.

Most genome-wide associated studies (GWAS) usually only examine single-nucleotide polymorphisms (SNPS), however, the largest DNA difference between individuals and cultivars are structural variants, including gene presence/absence variants. The most common source of structural variants is the activity of transposable elements (TEs), which are also responsible for large effect alleles. In addition, TEs contain transcription factor binding sites which can, if a TE inserts near or within genes, alter gene expression and rewire gene expression networks.

The aim was to assess the prevalence and impact of TE insertion polymorphisms (TIPs) by systematic analysis of 602 tomato genomes and transcriptomes.

b. Methodologies

Detection of TIPs

Illumina sequence was downloaded from three EBI-ENA projects for a total 602 tomato accessions, including wild tomatoes, *Solanum pimpinellifolium* (SP), early domesticated tomatoes, *S. lycopersicum cerasiforme* (SLC), and vintage and modern cultivated tomatoes, *S. lycopersicum lycopersicum* (SLL). Sequence was aligned to the tomato reference genome from SOL genomics (SLL cv. Heinz release SL2.5). TIPs were detected using SPLITREADER. Reads from TIPs were annotated by mapping to a TE library. TIPs were validated either by visual inspection of 600 randomly chosen TIPs using IGV or by PCR (2 TIPs from 22 accessions).

SNP calling and Linkage-disequilibrium analysis

The 602 aligned genomes were used to call SNPs using GATK. SNPs at the 8,760 positions genotyped by the SolCAP Infininum Chip SNP array were extracted. For each TIPs the pairwise r^2 was calculated between the TIP and 300 SNPs upstream and downstream. The percentage of TIP-SNPs and SNP-SNP comparisons that are in high LD were compared.

Genomic localisation of TIPs and genes

The position of TIPs and genes was represented by constructing a circos plot. The number of genes and TEs was calculated in 500-kb windows using Bedtools. Gene ontology was performed using AGRIGO.

Impact of TIPs on gene expression

RNA-sequence from (Zhu et al. 2018) was obtained and expression levels per gene was calculated by mapping reads onto the tomato reference genome using STAR. Counts were normalised using DESeq2. Normalised transcripts were compared between accession with and without TIPs.

Genome-wide association studies

Phenotypic information for 17 agronomic traits was extracted by web data scraping, chiefly from the World Tomato Society webpage. Metabolic and volatile data was obtained from previous publications (Tieman et al. 2017; Zhu et al. 2018) TIPs-GWAS was carried out using EMMAX.

Full-length cDNA nanopore sequencing

Total RNA was extracted from the ripe fruit of 400 accessions, converted to cDNA, which was then used for library preparation using the PCR Barcoding kit. The library was sequenced on the MinION MkID.

c) Main results and conclusions

Results

Most TIPs (84%) identified were Copia and Gypsy LTR retrotransposons. The chromosomal distribution of TIPs followed that of TE superfamily distributions, Gypsy TIPs near the centromere, while Copia TIPs were dispersed throughout genome. The richest diversity of TIPs was found in the SP group (wild tomatoes). Vintage and modern SLL had a more reduced TIP composition.

Genes with TE (mostly Copia) insertions were overrepresented in functions related to response to pathogens and other environmental stresses. Transcripts levels were compared between genes with at least one TIP within 1kb and those without any TIP within 1 kb. The presence of a TIPs was associated with both increased and decreased transcription and with changes in transcript features. Twenty percent (exonic) and 28% (intronic) of genic TIPs interfered with transcript elongation. Expression for immune and stress-responsive genes were particularly affected, for example a *COPIA* insert was identified in the exon of a CC-NB-LRR gene *Ph-3* which confers broad resistance to *Phytophthora infestans*. The insert was associated with transcript truncation which could therefore increase susceptibility to *P. infestans*.

TIPs with MAF > 1% and < 20% missing data were considered in GWAS for 17 important traits. Nine high-confidence loci were associated with important phenotypic traits (including fruit colour, leaf morphology and tomato flavour). The association with leaf morphology was much stronger for the TIP than any SNP and most TIP associations were not identified by SNPs.

Conclusions

TIPs are low-frequency variants rarely tagged by SNPs. They can be used in GWAS to identify gene/DNA changes associated with important traits.

2. *Hypothesis testing*

GWAS identified a TIP for 2-phenylethanol, which is a volatile that gives a pleasant flowery aroma to tomatoes. The TIP was a *COPIA* insertion in the intron of the gene *Solyc02g079490* which encodes for a protein with high similarity to a putative 2-phenylethanol Acyl-CoA transferase (*PPEAT*). *PPEAT* is involved in the esterification of 2-phenylethanol, which otherwise accumulates in fruits. The authors wished to test if levels of 2-phenylethanol were higher in tomato accessions with the intronic *COPIA* insertion.

I analysed the data for summary statistics based on the data supplied by the authors as supplementary data. The data provided showed the opposite of what the authors presented, the median level for the two samples was about the same (0.019 vs 0.013) and the mean was ten times lower in the sample with the *COPIA* insert rather than higher (0.105 vs 0.015). I emailed the corresponding author showing the summary statistics I had obtained and asked if there might have been a mistake with the data being uploaded. I did a *t*-test and tests for *t*-test assumptions in R based on the new data the authors sent me. I did not obtain the same *p*-value shown in the article, it was still highly significant, so I did not contact the authors again. However, the assumption of equal variances was not met based on the data I received, therefore a Welsh's *t*-test would have been more appropriate than the *t*-test reported. The authors did not address assumptions of the statistical test they performed either in the main article or in any of the supplementary data (that I could find).

a) relevant null and the research hypotheses being tested.

Null hypothesis: there is no statistically significant difference in mean 2-phenylethanol levels in cultivars carrying or not carrying the *COPIA* retrotransposon.

$$H_0: \mu_{COPIA+} = \mu_{COPIA-}$$

Research hypothesis: cultivars carrying the COPIA retrotransposon have statistically significant higher mean 2-phenylethanol levels that cultivars not carrying the *COPIA* retrotransposon

$$H_1: \mu_{COPIA+} > \mu_{COPIA-}$$

b) statistical test used and explain why it was appropriate.

A one-sided *t*-test was used to test if the mean 2-phenylethanol level was significantly higher in those accessions with the *COPIA* insertion in the *PPEAT* gene intron.

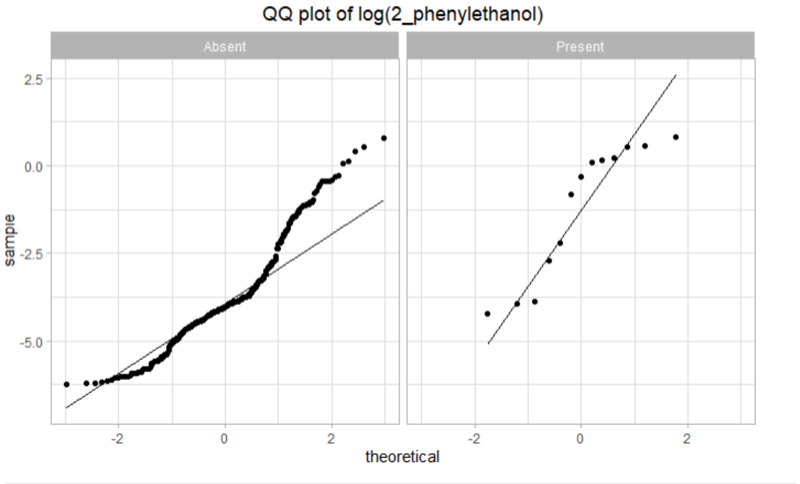
Why a t-test was appropriate

The data is in the form of 2 groups, i.e. accessions with and without the TE *COPIA* retrotransposon. There is one variable of continuous (numeric data), i.e. the level of 2-phenylethanol measured.

T-test assumptions met

The data is normally distributed: The data are (sort of) normally distributed. The article did not address any of the assumptions, but I did a QQ plot, shown below. The data are log (natural log) transformed, as shown in Figure 5d in the paper. The box plots shown in the article (Fig 5. d. below) also suggest that the data are normally distributed. The sample with the *COPIA* insert present is very small (13) and the

QQ plot is difficult to interpret. The box plot also shows that the data for *COPIA* insert present is skewed, the median line is not centred and the whiskers are shorter on one side than the other.



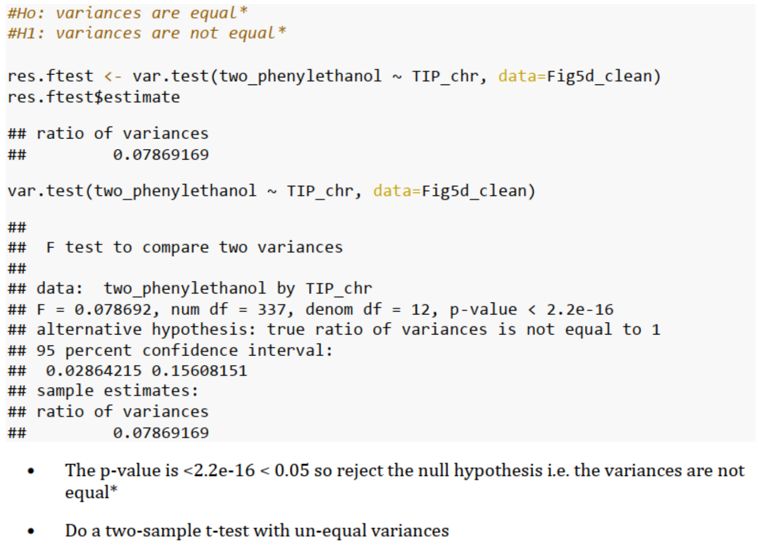
The samples are independent random samples. The samples are of all resequenced tomato genomes available with phenotypic data available. i.e. the samples weren't selected on some basis that might affect the 2-phenylethanol levels.

Type of t-test (unpaired, one-sided, Student's vs Welch's t-test)

The samples are unpaired: the samples in the 2 groups are unpaired.

One-sided vs two-sided: the question being asked was if the accessions with the *COPIA* insert had higher levels of 2-phenylethanol (and better flavour), so a one-sided *t*-test is appropriate.

The variances of the samples are equal: I did variance test in R (var.test), which suggests that the variances are not equal. A Welch's *t*-test would therefore be the most appropriate test to do.



```
#Ho: variances are equal*
#H1: variances are not equal*

res.ftest <- var.test(two_phenylethanol ~ TIP_chr, data=Fig5d_clean)
res.ftest$estimate

## ratio of variances
##         0.07869169

var.test(two_phenylethanol ~ TIP_chr, data=Fig5d_clean)

##
##  F test to compare two variances
##
## data:  two_phenylethanol by TIP_chr
## F = 0.078692, num df = 337, denom df = 12, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.02864215 0.15608151
## sample estimates:
## ratio of variances
##         0.07869169
```

- The p-value is <2.2e-16 < 0.05 so reject the null hypothesis i.e. the variances are not equal*
- Do a two-sample t-test with un-equal variances

c) *p*-value for the chosen test and draw conclusions.

The *p*-value is shown in Fig. 5 d. The p-value is 1.3 x 10^{-7}, which is much less than 0.05, we can therefore reject the null hypothesis and suggest that the mean level of 2-phenylethanol is significantly higher in tomato accessions with the *COPIA* insertion in the *PPEAT* gene intron. Incidently, the *COPIA* insertion is not found in modern tomatoes and the authors suggest that it has been selected against.

Provide an appropriate screen shot from the paper to support your answers

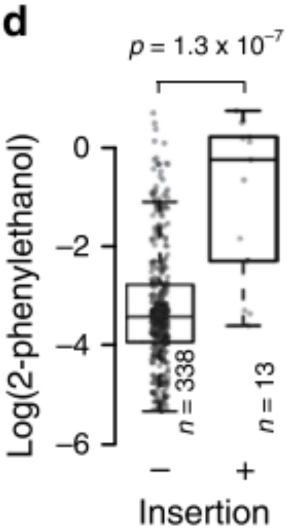


Fig 5. d. 2-phenylethanol levels in accessions carrying or not the intronic COPIA insertion. Statistical significance for differences was obtained using one-sided t test.

3. *Descriptive statistics*

Pick a graphical figure in the article and answer the following:

GWAS identified a TIP for leaf morphology. The TIP was identified as a *COPIA* insert in the gene *Solyc06g074910*, which the authors identify as the gene *BLi2*, which is a key regulator of leaf architecture (Busch et al. 2011). This TIP has been previously identified (Busch et al. 2011), which the authors use as validation that their GWAS method was a successful approach.

a) What is the figure displaying?

The bar char shows the percentage of accessions for each group (with or without the *COPIA* insertion) that have regular leaf or potato leaf morphology. The *p*-value for the Fischer exact test is shown above the bar chart. Above the *p*-value is shown the position of the *COPIA* insert in the gene intron. To the right of the bar char is line drawings of the leaf morphology.

b) What are the conclusions to the authors drawn from the figure?

The authors have run a Fisher’s exact test on the data. Fisher’s exact test is used to determine if there is a significant relationship between two categorical variables. It is used when the sample is small instead of a χ^2 test.

Null hypothesis: There is no relationship between the two variables, in this case, between carrying the *COPIA* insertion or not and leaf morphology (regular or potato)

Research hypothesis: There is a relationship between carrying the *COPIA* insertion or not and leaf morphology (regular or potato).

The *p*-value is less than 0.05, so we can suggest that there is a relationship between the *COPIA* insertion in the BLi2 gene and leaf morphology, accessions with the *COPIA* insertion are more likely to have potato leaf. This suggests that the *COPIA* insertion modifies the gene transcript.

Provide an appropriate screen shot from the paper to support your answers.

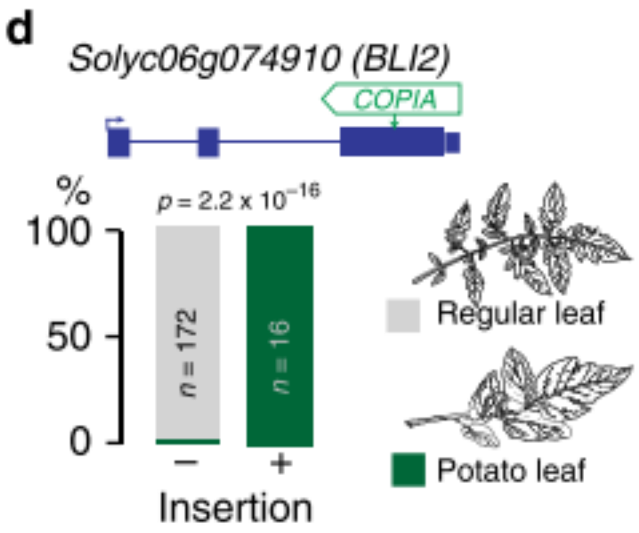


Fig 3. d. Leaf morphology of accessions carrying or lacking a *COPIA* insertion within *BLI2*. Statistical significance for differences was obtained using two-sided Fisher test.

4. References

Busch, Bernhard L., Gregor Schmitz, Susanne Rossmann, Florence Piron, Jia Ding, Abdelhafid Bendahmane, and Klaus Theres. 2011. “Shoot Branching and Leaf Dissection in Tomato Are Regulated by Homologous Gene Modules.” *The Plant Cell* 23 (10): 3595–3609. https://doi.org/10.1105/tpc.111.087981.

Tieman, Denise, Guangtao Zhu, Marcio F.R. Resende, Tao Lin, Cuong Nguyen, Dawn Bies, Jose Luis Rambla, et al. 2017. “A Chemical Genetic Roadmap to Improved Tomato Flavor.” *Science* 355 (6323): 391–94. https://doi.org/10.1126/science.aal1556.

Zhu, Guangtao, Shouchuang Wang, Zejun Huang, Shuaibin Zhang, Qinggang Liao, Chunzhi Zhang, Tao Lin, et al. 2018. “Rewiring of the Fruit Metabolome in Tomato Breeding.” *Cell* 172 (1–2): 249-261.e12. https://doi.org/10.1016/j.cell.2017.12.019.