UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE

**DEPARTMENT OF GENETIC MEDICINE & DEVELOPMENT**

**CMU** - Rue Michel Servet 1 / CH-1211 Genève 4
phone. +41(0)22 379 54 83- Fax +41(0)22 379 57 06
www.unige.ch/medecine

***Professor Emmanouil (Manolis) Dermitzakis***
emmanouil.dermitzakis@unige.ch

RE: Nature manuscript 2012-12-15883

March 4, 2013

Dear Magdalena,

We would like to thank you for giving us the possibility to revise the manuscript and for the valuable feedback from you as well as from the reviewers that was of great help in the revision process. As a result, we feel that the manuscript has improved substantially, and we hope that it is now suitable for publication in *Nature*.

Below, we provide a detailed response to the specific points raised by the reviewers. However, there were three major concerns that were raised: (1) the novelty value of the study, (2) technical quality and replicability of RNAseq data processing, and (3) the manuscript being too dense. We have addressed all these points in the revision, and would like to respond to these points with the following general statements:

1) We must disagree with the Reviewers 1 and 3 that question the novelty of our conclusions. Even though this is not the first population-level RNAseq dataset, we claim that this is actually the first really mature RNAseq study of multiple human populations with both mRNA and small RNA data. The high quality and quantity of raw data as well as developed analysis methods allows us to get beyond technical issues to really discover novel biological phenomena. Many of our results are novel and far from obvious: e.g. the special role of splicing in population differences, the abundance of regulatory variation that was never discovered with even large array studies, the independence of genetic variants affecting transcript levels and transcript usage, miRNA-mRNA interactions in populations, added analysis genetic variants underlying allele-specific expression, and the resolution of the analysis of eQTL variants with respect to recently reported comprehensive functional annotations. Furthermore, our characterization of causal variants and validation of loss-of-function variants contributes directly to some of the hottest topics in the field and has real practical implications for medical genetics studies, and serves as a model for combining other types of phenotypic data with genome sequencing data. Even though we chose not to focus on methods development in this paper, we describe novel approaches especially in allele-specific transcription analyses.
2) Reviewers 2 and 3 expressed a general concern about the robustness of the results regarding the complex RNAseq data processing steps, which is a reasonable general concern in all studies with big data. However, at the same time the quality of the analysis was complimented, and the few specific concerns we address below. We feel that it is unfair to our study to judge it based on general technical concerns that potentially might exist in the absence of specific indications of any. Our methodological choices are naturally supported by much deeper analysis than what was included in the Supplement due to length constraints, but we have made additional effort in the revised version to improve documentation of data processing in the Supplement. While it is impossible to share every line of our data processing scripts, we have made substantial effort to be transparent and share data, resources and documentation. This will allow our analyses to be replicated by others, which was a major concern of Reviewer 2 who mistakenly thought that data access is password protected. Specifically, the raw data has been freely accessible since November 2012, additional data files will be accessible after publication similarly to the project wiki with additional analysis results and discussion (see links below).
3) The reviewers as well as the Editor expressed concern that the manuscript might be too dense, with analysis of a wide spectrum of topics that are difficult to cover in adequate depth, rather than focusing on

a few key points. We agree with this to some extent, and have omitted the following topics from the manuscript in order to improve its focus:

- Gene discovery as a function of sequenced individuals (formerly pp. 4-5, Fig S10)
- Subtle splicing (formerly pp. 5, Fig S13)
- Fusion transcripts (formerly pp.5-6, Fig S14, Table S4)
- RNA editing (formerly pp. 8, Fig S20)

Additionally, we strengthen our most important theme of genetic regulatory variants by adding the following novel analyses:

- mapping genetic regulatory variants and replicating eQTLs using allele-specific expression data (Fig. 3c, S31-S33, pp. 11-12 in the main text and section 12.3 and 12.4 in Supplementary Methods)
- as per the advise of reviewers we have explored the functional validation of our eQTL putative causal variants in two published datasets, finding strong enrichment of allele-specific binding of CTCF, as well as strong enrichment in DnaseI QTLs (dsQTLs) (Fig. S21, pp. 10 in the main text, section 11.5 in Supplementary Methods).

With these changes, the manuscript has the following key themes:

- Quantitative versus qualitative variation in human populations: analyses comparing expression levels and transcript structure in (1) the transcriptome alone, (2) QTLs, and (3) allele-specific effects
- Micro-RNA variation in human populations and its effect on the transcriptome
- Functional mapping of putative causal regulatory variants using data from eQTLs, allele-specific expression, and loss-of-function variants, as well as characterizing the properties of these variants

We would like to emphasize that this is not a narrow hypothesis-based study addressing a single important question, but rather a reference study of the use of transcriptome sequencing and genome sequencing to understand genome function variability. We are confident that the biological insights, the methodological and technical advances, and the easily accessible data will have a large impact in functional and medical genomics in a similar manner to the 1000 genomes and the HapMap papers. In line with this, the main paper will be followed by a series of companion papers that analyze various topics in further depth: e.g. technical aspects of RNAseq (manuscript included in the submission and under review in Nature Biotechnology), population variation in splicing, splicing QTL mapping methods, loss-of-function analysis, population genetics of regulatory variation, and noncoding RNAs. We preferred to prepare the main paper first rather than keep it waiting for the companions, so that the community will have access to the data and the main results without delay.

We would like to note that in addition to the changes above, we have made a small improvement in the normalization of exon and repeat quantifications, resulting in less than 1% changes in total eQTL numbers and consistent results from other downstream analyses.

Our aim has been to make data and documentation as accessible as possible. The main links to the online resources for data access and documentation are the following:

- Project website (no password required): http://www.geuvadis.org/web/geuvadis/RNAseq-data-release
- ArrayExpress (the main portal for data access, no password required, accessions E-GEUV-1, E-GEUV-2, E-GEUV-3): http://www.ebi.ac.uk/arrayexpress/browse.html?keywords=E-GEUV-*
- The Geuvadis Data Browser: (no password required for visualization. Regional data view has been implemented but is inactivated until publication): http://wwwdev.ebi.ac.uk/Tools/geuvadis-das/
- Analysis group wiki (username: Geuvadis.Visitor, password: geuvadiswiki): geuvadiswiki.crg.es

We hope that the revised version of the manuscript is now satisfactory and suitable for publication.

Sincerely,

Tuuli Lappalainen
Emmanouil Dermitzakis

**SPECIFIC RESPONSE:**

**REVIEWER #1**

**The authors report an RNA-seq study of 465 cell lines from the HapMap/1000 Genomes collection. This is the largest RNA-seq study to date (although comparable in size to published microarray studies in many of the same samples). The paper describes the functional context of eQTL variants, splicing variants and so forth. Overall, the paper is well written and clearly presented.**

**That said, unfortunately, I feel that we do not learn enough here to justify an article in Nature. While this data set is bigger than earlier RNA-seq data sets, most of the take-home points in the paper have been considered elsewhere, especially in papers by Lee et al (2009), Montgomery et al (2010), Pickrell et al (2010), MacArthur et al (2012) and Gaffney et al (2012). I think that the paper would be better suited to a specialty journal.**

RE: We discuss the general novelty issues of our study on page 1 of this document. Many of our results are completely novel, and we feel that early indications of some of the results in previous studies should not preclude high impact publication of the definitive result.

Changes in the manuscript: We have added discussion of the novelty value of our data and results throughout the manuscript, especially in introduction and discussion

**Figure 1c: "transcript structure variation appears to contribute disproportionately to continental differences, suggesting a special role of splicing in human adaptation." The authors report an enormous difference between within-continent comparisons and between-continent comparisons. It's hard to conceive of a plausible model--either neutral or adaptive--that is consistent with this enormous variation in the relative fractions of gene-level vs transcript level changes, leading me to think that this is most likely artifactual. Perhaps this observation may be due to differences in power at the two levels?**

RE: In this analysis, we compare the relative contributions of expression level differences and transcript usage differences between populations, and we observe that is radically different for intercontinental population pairs. We have several reasons to be convinced that our result is driven by real biology. Figure 3c is based on analysis of differentially expressed genes/transcripts, and this result is replicated using two other methods (quantitative analysis in Fig. S11d, and now added DEXSeq for quantifying differential splicing in Fig S11c). Furthermore, two recent papers in *Science* (PMID 23258891 and 23258890) describe adaptive evolution through splicing changes in interspecies variation, which is completely independent data and analysis but perfectly consistent with our result from human populations. We do not think that the difference YRI-EUR vs EUR-EUR population pairs is too large to believe – the reason why it hasn't been found before is that this is the first data set where this can be properly analyzed. We rigorously explored the data and analysis and could not find any technical reason why YRI-EUR population pair comparisons would be systematically different from EUR-EUR population pairs: the sample sizes of different populations are similar, the sample processing for sequencing was randomized, and the effect is relatively consistent across cell line collections of different age.

Changes in the manuscript: We have included a third analysis method to support the result (DEXSeq analysis in Fig S11c), included additional discussion in the legend of Fig. S11, added references and revised the main text for additional clarity and caution (pp. 5).

**Several of the figure panels are relatively uninformative, including most of figure 1 as well as figure 3b. The analysis of NMD displayed in figure 4 is a simple update of a corresponding figure in MacArthur et al.**

RE: These figures show distributions of different phenomena (rather than barplot-type plots that capture numerical summaries of the data), and we feel that together with the summary statistics given in the text, a visualization of the shape of these distributions is important and informative. The loss-of-function analysis in Figure 4 shows inter-individual variation in NMD unlike MacArthur et al., and the quantification of NMD described in the text is more advanced and definitive with many more sites interrogated.

**REVIEWER #2**

**This study represents some of the very best in contemporary genomics research, but also one of its darker aspects. The best is that you cannot fail to be impressed by the magnitude of the study, quality of the analysis, comprehensiveness of treatment, and attention to ensuring that the data becomes a crucial community resource. The dark is that it is really 20 papers condensed into one, with some very interesting topics given just a sentence with a deeply buried supplementary figure (eg RNA editing SNPs), but more troubling is that it also means that several of the major findings are not given the depth of coverage to convince at least this reviewer beyond reasonable doubt.**

RE: We would like to thank the reviewer for the enthusiastic support and for constructive criticism, which we are happy to respond to. After discussions with the Editor, we have dropped several less important analyses (see page 2 of this document) in order to improve the focus of the paper and make it less dense to read. We would like to emphasize that this is indeed the main paper of a consortium that will publish additional papers based on this dataset; we provide the pdfs of the companion papers focused on QC and splicing with this submission, and list the upcoming companion papers in the project webpage for everyone to see.

Changes in the manuscript:

We have removed the following analyses:

- Gene discovery as a function of sequenced individuals (formerly pp. 4-5, Fig S10)
- Subtle splicing (formerly pp. 5, Fig S13)
- Fusion transcripts (formerly pp.5-6, Fig S14, Table S4)
- RNA editing (formerly pp. 8, Fig S20)

Additionally, we have strengthened our most important theme of genetic regulatory variantion by adding an analysis of ASE-based mapping of regulatory variants (pp. 11-12 in the main text).

**With over 50 pages of supplementary material, it is unlikely that my review comments will jibe with those of other reviewers, and I have to confess that I am concerned that the most appropriate reviewers should be post-docs or students who know their way around the details of the bioinformatics tools. I had some of my people read it and our consensus was that there is really no way of being sure of detailed claims without replicating the studies with each group's favorite pipelines. No doubt the big picture is largely correct, but my major comment is that data access is going to be crucial.**

We discuss the questions of data access and robustness of the results above (page 1-2) as well as in the next response, and we indicate this better also in the manuscript. We encourage replication and other additional analyses by other researchers and we would be happy to provide any information or intermediate data files for the reviewer to test and execute some of our analysis. However, we believe that the review process in a journal should rely on whether the methodologies described are reasonable and not whether they can be replicated during the review process, which in most cases is unrealistic.

Changes in the manuscript: We have added independent analyses to replicate our key findings, such as ASE-based mapping of regulatory variants that strongly supports our eQTL discoveries (Fig. S32-S33, Fig. 3c), and included more discussion on robustness of the results throughout the Supplementary Material, and better documentation of data access (see the next response). Additionally, for additional transparency, we have improved the content on the project website and provide the reviewers pre-publication access to the project wiki, which will be fully open once the paper is published (see page 2 of this document).

**The authors have promised to make the data available through ArrayExpress and the European Nucleotide Archive, but the latter is password protected. Please clarify what the nature of the Data Access Committee will be, and what steps potential whole genome analysts will need to take to gain access.**

RE: We are extremely surprised if the reviewer has been prompted for a password when trying to access the data – this should not happen, the raw data and bam files have been freely accessible to everyone since November 2012 through ArrayExpress, without any passwords or data access committees. If the Reviewer has any problems accessing the data, please let us know via the Editor in order to identify the problem.

Changes in the manuscript: We clarify the open access policy better in the main text on pp. 4 and in the Data Access section (pp. 15). and provide a data access schema in Fig. S36 with a detailed legend. We have also updated the project web page to include better documentation.

**The second is the thousands of candidate gene surveyors who are more likely to simply want access to the Geuvadis server, which is a terrific resource, but I think it is essential that users also gain access to the single gene-relevant data upon which the graphical output is based. Please clarify whether users will be able to download the sequences (genome, RNASeq, miRNASeq (or at least inferred abundance in tab format) for single genes. Graphical output is nice, but since it only hints at the causal variant, sophisticated users will want the actual data without having to access the entire dataset.**

RE: We agree with the reviewer that this is a valuable element for the Geuvadis server, and we have enabled access to quantifications and QTLs by region. We believe this will be adequate for most users who do not wish to analyze the full files. Bam files can be visualized in the Ebsembl browser but unfortunately it has no regional splicing function, and since our project was not funded to produce a full-blown browser, it is beyond our resources to implement this.

Changes: We have implemented a regional data access function in the Geuvadis Data Browser. However, since we do not wish to share the QTL and quantification data before publication, this feature will be activated later.

**1. That most of the variation in exon-level read counts is due to transcript ratios rather than whole gene expression, and that a significant proportion of this is among populations. This conclusion is probably quite sensitive to the metric used to assess the relative proportions of tr and ge. Readers have to dig all the way through the Suppl Methods to find that it follows Gonzales-Porta et al (ref 46, not 47), the philosophy of which needs to be explained briefly to readers in the text. My concern is that it is sensitive to the contribution of relatively rare classes, and that if there was an exclusion of rarer transcript types the inference would be quite different.**

RE: We have made effort to clarify the methodology in the main text; however, keeping the length limitations in mind we believe that too much mathematical detail of a previously published method would not serve the majority of the readers. We have analyzed the contribution of expression level to this analysis and include the results in Fig. S11a, and find the signal to be consistent – lowly expressed transcripts show slightly less contribution by splicing variation, if anything, likely due to lack of resolution.

Changes in the manuscript: We have improved the writing in pp 5, and added results in Fig S11a to show that the estimates are robust to gene expression levels.

**What may be very confusing to readers is the juxtaposition of this inference with the conclusion that there are several thousand eQTL for gene abundance but only a few hundred for transcript ratio abundance. This is explained away as a power issue, but needs to be dissected more deeply: is it that power is lower because the abundance of contributing classes is low?**

RE: Read counts of exon levels as well as whole gene levels are observed quantities that are well measured given the depth of the read data. On the other hand, transcript ratios are estimated based on models of read pair distribution and known annotation and this estimation is noisier, and thus eQTL analysis of exons and whole genes is more powerful than in transcript ratios. It is also possible that that transcript usage is truly less affected by genetic variants than expression levels, but given these technical constraints, our data does not permit such a conclusion.

Changes in the manuscript: We have clarified the text on pp 7.

**Another aspect relevant here is the statement (which I completely missed on first reading) that most of the transcript ratio variance is not exon-skipping, but in 5' and 3' terminal exons, implying that the tr variation does not contribute greatly to protein variation.**

RE: We agree with the reviewer that this is an important point.

Changes in the manuscript: We have emphasized this point on pp 8.

**2. That RNA-Seq helps in the annotation of causal variants.**
**This section could also be presented more clearly. It starts with the observation that there is enrichment for ENCODE-type features that is particularly notable for the "best" eQTL, then builds into a somewhat ad hoc estimate that 57% of the best eQTL are the true eSNP based on a somewhat arbitrary NLP>1.5 cutoff (the plateau is more like a gentle slope), and concludes that there is enrichment for GWAS hits. I think that Figure 2D is really the critical one, and suggest turning the**

**discussion around, beginning by asking the question directly "how does RNA-Seq improve disease variant annotation", presenting a case study, and leading the reader through the process. The fact is that in half the cases the "best" eSNP in a statistical sense is not going to be the causal one, which is to be expected (the strongest statistical SNP is likely to be a function of methodology and sampling anyway). New Bayesian and other methods are being developed that will help refine likely causal variants further, which places a premium on full data access as argued above. My point is that enrichments of enrichments are not overly convincing, so perhaps better just to walk readers through how to use the resource rather than generate difficult-to-interpret figures like 2A.**

RE: We have added results of more conservative threshold choices in the manuscript and clarified that our estimates are approximations. If we understand correctly, the reviewer shares our view that developing novel sophisticated causal variant modeling methods is beyond the scope of this paper, but we agree that our data will be of great importance for this purpose in the future. Regarding the structure of this section, we felt that starting with GWAS might be confusing since that is not the focus of the manuscript in general, and thus we decided to keep the order of the paragraph as it was.

Changes in the manuscript: We have added more conservative estimates of causal variant proportions and softened the statements in this paragraph (pp. 10).

**3. That there is substantial variation due to relatively rare variants that impact transcript ratios in particular.**
**This is very believable, but I am just nervous about it because RNASeq analysis is subject to so many biases. This is not really a criticism, just a statement. In general the authors have been very open about their analyses, use state of the art methods, and are very careful. But, different methods of normalization, alignment, transcript inference, and association testing could yield different results. I am not asking for reanalysis, but do feel that a cautionary statement and call for verification with other methods is appropriate. For example, PEER is excellent for cis-eQTL analysis, but I can well imagine it upsets the trans-miRNA associations and maybe affects the rare exon variant distribution. Similarly, the Flux Capacitor approach is just one way to infer transcript structure from paired-end reads - different graph theoretic algorithms (eg Cufflinks) will likely give different results. Of course the authors have the right to choose their method, and others can replicate it, but less familiar readers should have a sense of how robust the major biological conclusion is to methodology. Just one example of why one might be skeptical is in Fig 1C: it makes little sense that the tr:ge percentages are so different for the GBR:TSI and FIN:TSI comparisons than the CEU:TSI and CEU:FIN ones.**

RE: We agree with the reviewer that it is difficult to evaluate the robustness of the results when data processing is becoming increasingly complex; see our response on page 1 of this document. We also thank the reviewer for understanding that in practice, in addition to the measures that we already undertake, there is not much else that we can do at the moment. We can assure the reviewer that the analyses and data processing methods have been evaluated to a much greater depth than it is possible to present even in the Supplement due to length constraints. We have made substantial effort to make our analysis transparent, be conservative, evaluate the robustness of our methodology, and provide diagnostics of the most important methods (e.g. transcript quantification in Fig S10 and PEER in Fig S7). Furthermore, many results are supported by two or more methods (e.g. Fig 1c, Fig S12 now with additional results, and ASE-based replication estimates of eQTLs), and we refrain from biological interpretation of results when we are not confident that the result is not affected by technical biases that we were able to think of. The differences between European population pairs in Fig 1c are a good example of this as it may be affected by cell line age – however the main conclusion that we reported, the difference between YRI and European populations, appears reasonably robust to this (see also response to Reviewer 1 on page 2).

Change in the manuscript: Additional discussion of robustness of the results throughout the Supplementary Material.

**Some minor points:**
**The inference that the differences between YRI and EUR expression profiles is due to genetic divergence should be softened, given that batch effects of LCL production are known to cause expression divergence (see the Cheung/Storey discussion a few years back)**

RE: Change in the manuscript: We have softened the statement and discuss cell line batch effects in the legend of Fig. S11. See also response to Reviewer 1 on Page 3 of this document.

**Regarding the lab effect, Fig SF6 shows that it is actually quite substantial, and much stronger than population effects in the raw data. That PEER removes this is not surprising, and is good news for the analysis, but it does not necessarily mean it properly adjusts for it. I feel the authors make too strong a claim that lab differences are negligible.**

RE: We have analyzed and quantified lab effects further in the accompanying paper ('t Hoen et al.) that is included in the submission. The biological differences between our samples are very small and thus e.g. in Fig S6 the lab effect being the strongest trend in the data still does not imply that it is a very strong effect. Importantly, clustering methods such as MDS are visualization methods with heavy compression of the data, rather than proper quantitative analysis. In our opinion, the best proof of lab effects being smaller than biological variation is the numerical analysis of the correlation values of our replicate samples in Fig. 1a.

**The section on miRNA-mRNA covariance requires more scholarship - there is quite a literature claiming feedback etc in cancer samples to explain relatively low correspondence between suspected targets and miRNAs. Please add some context and references.**

RE: After careful consideration of the reviewer's comments, we decided to focus on more global phenomena that link our findings together with previous knowledge and hypothesis of miRNA function, rather than on individual examples as before. We have also added more references to earlier papers about miRNA-mRNA interactions.

Changes in manuscript: We have rewritten the miRNA-mRNA section of the manuscript (pp. 6).

**Middle of p7 it is claimed that gene eQTL and tr eQTL are expected to be orthogonal, contrary to the observed 45% overlap. It is not clear to me why they are expected to be orthogonal, particularly if most of the tr effects are in the 3' and 5' ends**

RE: We agree that the wording of that particular sentence was not optimal. What we meant by "orthogonal" was that our measurement of gene expression levels and transcript ratios are independent – a biological change in only one does not change the statistical measure of the other, thus any correlation is likely to be biological rather than technical.

Changes in the manuscript: We have rewritten these sentences to improve clarity (pp. 7) and explained this further in the legend of Fig. S15).

**In discussing the RTS analysis, we need an explicit definition of "best" eQTL. Is this just the one with the smallest p-value, or are you referring to the best one conditioned on other evidence such as TFBS, DHS, histone proximity?**

RE: The best eQTL is just the one with the lowest p-value, as we defined on p. 8.

**One potential source of technical error contributing to Fig 3B is read alignment artifacts. Were QC checks performed to ensure that all ASE and ASTR tests were done on SNPs that are present in both the RNASeq and WGS data for that individual? Ideally the alignments would be performed against the person's individual genome, not RefSeq - does this have an impact on rare ASE annotation?**

RE: We fully agree that allelic mapping bias is an important issue in RNAseq. We have actually done a substantial amount of work to address this issue, and we now include these analyses in the manuscript. Based on simulated mapping bias of all 1000 Genomes variants, we exclude about 12% of variants from ASE analysis that we consider unreliable due to mapping issues. In addition, when personalized mapping was performed in a small subset of samples we observed little to no difference in ASE estimates vs. mapping to the reference genome and correction based on simulations. The SNP genotype data accuracy is much higher for the 1000 Genomes Phase 1 data than for the pilot so genotyping errors are not a major issue (as seen in Fig S28), but to account for any remaining errors, we use only sites where both alleles are observed in RNAseq data. Furthermore, we are preparing a separate paper addressing the effect of mapping bias on eQTL discovery (Panousis N, Gutierrez-Arcelus M, Dermitzakis ET, Lappalainen T in preparation), and we have found that the effect is negligible and causes false eQTL discoveries only in about a dozen genes out of thousands. We prefer not to include those results in this paper, but the figure below shows the p-values before and after filtering reads that may be affected by mapping bias, demonstrating that this is not major concern in eQTL analysis of RNA-seq data.

Change in the manuscript: We have added section 4.1 in Supplementary Methods and Fig. S26 of the analysis of allelic mapping bias.

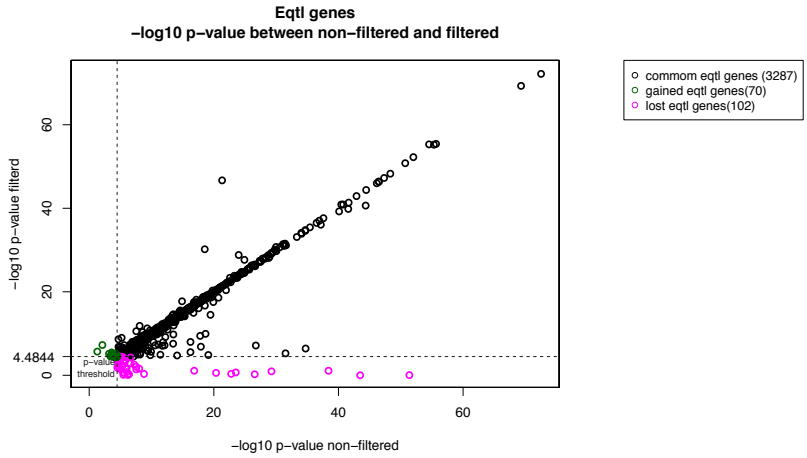


**Fig 3c shows that ASTS tend to be enriched in introns and 3' UTR whereas ASE are relatively enriched in splicing regions, contrary to what I would expect. Please discuss.**

RE: We first found this surprising as well, but this pattern actually makes sense when considering the fact that we are only assessing the reads that cover a particular site. At least 16 reads of coverage over fully intronic sites is rather unusual, and can be due to rare intron retention effects or unannotated exons, both with unusual splicing easily leading to an ASTS signal. The ASTS signal in the 3' end is less surprising knowing the widespread variation in 3' ends of transcripts. Splice site variants are an interesting case; for example, let's consider a nonreference allele that decreases splicing efficiency by 60%. In this situation, 60% of nonref transcripts would skip the exon or undergo NMD, leading to a strong ASE signal over the splice site – however, the remaining 40% of the reads over the splice site that we would assess for ASTS would have the normal splicing pattern, without ASTS over this site.

Changes in the manuscript: We have discussed the interpretation of this figure (now moved to the Supplement, Fig. S30) and its legend.

**What is the average read depth of the 1000G data?**

RE: We assume that the reviewer is referring to the genome sequence data. For the low-coverage sequencing of the whole genome the average depth is 5x, and for exomes it is 80x, as specified in the 1000 Genomes Phase 1 paper.

**Which FDR method was used throughout - some figures refer to Pi0 estimates, suggesting q-value analysis, but I suspect Benjamini-Hochberg FDR has been used.**

RE: In all QTL analyses we used permutations, in the miRNA-mRNA correlations Holm's procedure was used, and differential expression analysis uses Bonjamini-Hochberg, as defined either in the method description or in Supplementary Methods. The Pi0 statistics e.g. in Figure 3a is used as a quantification of sharing of the signal and not to estimate FDR per se.

**This is a lot of critique, but there is a lot to review. I just want to close by complementing the authors, and emphasizing that the criticisms are intended to assist in accessibility for a wide audience. In writing this review, it become apparent to me that my major recommendation for the paper is to seriously consider concentrating on just the few main biological issues and provide more evidence and analysis, while dropping the references to half a dozen other topics that really belong in other papers. I don't feel strongly about this - it is all interesting - but there is some frustration with strong claims that are probably true but could be supported with a sense of robustness to methodology for a subset of the data.**

RE: We would like to thank the reviewer for the compliment and constructive criticism that was of great help in improving the manuscript. We have made substantial changes in the paper to make it more focused and leave more space for discussing the core messages.

Changes in manuscript: See previous response on pp. 4 of this document.

**REVIEWER #3:**

**This paper provides a comprehensive analysis of human genetic variability. The authors acquired deep coverage RNA-Seq data from 465 individual samples covering 5 populations of the 1000 Genomes project and carried out intensive transcriptome and genotype comparisons within and across populations. Based on the reported statistics of the individual samples, libraries, and biological replicates, we can see the overall data quality is very good. With these data, authors explored transcription variation, eQTL and trQTL, loss of function variants and identified some new loci as the causalities of phenotypes/diseases. Most methodologies adopted in this work were all recently developed. These methods tackled many known biases. However, there were still questions in need of clarifications.**
**Overall Verdict: Major Revisions are needed**

**My overall feeling on the manuscript is that this is certainly an interesting subject area connecting functional data to genome variation, and the paper represents a substantial dataset. I was quite excited to read it. I do believe it has merit. However I think it requires large revisions before being acceptable. In particular the authors have to address the question of what novelty their analysis brings to this and they have to address a lot of technical RNA-seq processing issues, which I find fairly sloppily dealt with here. Again these are areas that I think can be fixed but will require substantial recalculations. It is somewhat disappointing that the authors did not do this initially.**

**The fundamental issue with this paper as I see it is, what is new? Obviously one has about 10 times as much sequence RNA data as previously in the path-breaking Pickrell et al. and Montgomery et al. papers. A lot of data has been generated but what new results do we see? The authors have to make this clear. It cannot just be simply that we find more eQTLs. One really has to understand how all this data gives rise to new ideas. I think the authors need to spend a considerable amount of text talking about how their current work qualitatively, not quantitatively, goes beyond previously published analyses, if, indeed, it actually does, with clear demonstration that the earlier results would not be possible with the smaller datasets.**

RE: See the response on novelty value on page 1 of this document. While we understand the reviewer's point, we believe we have added substantial biological insight and analyze the data far beyond merely reporting the numbers of eQTL discoveries.

Changes in the manuscript: We have added discussion of the novelty value of our data and results throughout the manuscript, especially in introduction and discussion.

**More to the point, the results are consistent with our understanding and expectations from human genetic variability studies. However, just because most results could be anticipated, it seems there is lack of thrilling discoveries from this wonderful data set. Also, some more work such as following up experiments to show the molecular/biochemical mechanisms of the discovered causal regulatory variants could add great value to this work.**

RE: While it is of great value to perform follow-up experiments, it is out of the scope of this study. Testing biochemical mechanisms requires a completely different set up, and it would be difficult to achieve reliable validation estimates beyond individual case examples that in our opinion would be of limited value. However, we have now used existing data sets from McDaniell et al. (Science 2010) and Degner et al. (Nature 2012) to investigate allelic effects of our eQTLs in CTCF binding as well as enrichment in DNase1 hypersensitivity QTLs and have observed strong enrichment in allelic binding effects (CTCF) and overlap with dsQTLs (DNaseI). Furthermore, the samples are available from Coriell and used by many laboratories, so in the future there will hopefully be additional data sets that can be linked to ours.

Change in manuscript: We have added analysis of allele-specific binding and dsQTL overlap (Fig. S21, pp. 10 in the main text, section 11.5 in Supplementary Methods).

**The paper discusses the relationship between variants and total gene expression level, the expression level of exons and also of transcripts. Obviously the biologically relevant unit here is that of transcripts and one would expect to get the strongest correlations and the clearer signals in relation to that. In the paper the converse is true, far fewer eQTLs are found for transcripts than for exons and genes. Perhaps this is because of deficiencies in transcript structure in some of the RNA-seq processing method used. This needs to be explored in considerably greater statistical detail to understand why one is getting such a counterintuitive and to some degree obviously wrong result.**

RE: There seems to be a misunderstanding here – the trQTLs are not for transcript expression *levels* but for transcript *ratios* with the aim to find variants that affect the choice of transcript that is expressed rather than the level of transcriptional activity. These are two different quantitative traits, and thus the results are not expected to be similar. See also response to Reviewer 2 on page 5 discussing why the numbers of eQTLs and transcript ratio eQTLs may differ.

Changes in the manuscript: We have clarified the text on pp 7.

**Also as I am sure the authors well know, there is a huge controversy in the world of RNA editing, a lot of it requiring substantial validation. The authors should allude to this controversy and either improve the content of the manuscript technically or remove this section. Their discussion of editing could have been improved by more validation related to these results.**

RE: The difficulty of reliable calling of editing sites is the reason why we analyzed individual differences in *known* edited sites, which is much less sensitive to error than determining if a particular site is edited in general or not. However, in order to obtain better focus of the manuscript and after discussions with the Editor, we have omitted this analysis from this manuscript.

Changes in the manuscript: The RNA editing analysis has been omitted

**3. I was unimpressed with the consideration of read mapping. The whole point of the paper is to look at variants in relation to RNA-seq and one's using essentially a mapping strategy does not take into account these variants, mapping everything directly onto the reference genome. There are many additional papers now that are starting to take into account this type of thing, either in relation to local assembly or other strategies. I really think that in this type of paper in 2013 one should see a more developed mapping strategy being utilized. This becomes particularly important for editing and for allele analysis as well as for the eQTLs. See http://www.ncbi.nlm.nih.gov/pubmed/19808877 and http://www.ncbi.nlm.nih.gov/pubmed/21935354 .**

RE: See response to Reviewer 2 on page 7-8 of this document.

Changes in the manuscript: We have added section 4.1 in Supplementary Methods and Fig. S26 of the analysis of allelic mapping bias.

**4. Another issue was the linear regression for quantifying eQTLs. There have been several papers discussing various algorithms to detect eQTLs (e.g. http://www.ncbi.nlm.nih.gov/pubmed/19303049), and the papers claimed significant inconsistencies of eQTLs resulting from different algorithms. Some suggested the linear regression model might overestimate the amount of eQTL (such as in the same paper). It is thus necessary to further examine eQTLs using some other models.**

RE: We agree that linear regression is not the most sophisticated eQTL method, and is known to generate false positives when outliers are present. However, linear models are useful when one wants to use covariates. In addition, transformation of each individual phenotypic distribution (exons, genes, transcripts etc) to a standard normal and performing permutations makes us immune to outlier problems. In our long experience with eQTL analysis, the inconsistencies described by the reviewer, while they exist, they do not necessarily suggest incorrect analysis but rather differences in the type of discoveries. In fact, there is a lot of debate as to whether more complex models are better or relevant to eQTL analysis than the more traditional SNP by SNP tests and we did not feel that our paper is the place to address this debate.

**5. A final technical question is the normalization of mRNA/miRNA. A recent paper showed that c-Myc had a huge impact on the total amount of RNA, and if this effect was ignored, it was very likely to draw some unreliable conclusions. (http://www.ncbi.nlm.nih.gov/pubmed/23101621). We do not know the direct transcription variations of c-Myc across the 465 individuals. But as shown in the supplementary**

**results (Figure S22, Figure S24), the gene c-Myc was among the identified eQTLs as one of outstanding example. If c-Myc transcription variation was indeed large, the authors will have to take this aspect into consideration and revise the results accordingly.**

RE: We agree with the reviewer that one should always be careful with data normalizations and evaluate its impact on the results. However, our focus on cis-eQTLs makes our analysis much less vulnerable to such effects that would indeed be a serious concern for trans-eQTL analysis or transcriptome profiling. The annotation overlap figures (now Fig. 2a,S18, S19) do not actually indicate that c-Myc has or is an eQTL, but that c-Myc binding sites are enriched for cis-eQTLs. This does not imply that c-Myc levels in trans necessarily drive these eQTLs in cis. More generally, we are aware that both known and unknown factors may be influencing global levels of expression but it is hard and statistically risky to correct for a large numbers of covariates. Instead, our approach to correct for PEER factors is expected to eliminate the effect of the global factors with strong effects. We further verified that PEER is not correcting out biological effects in the miRNA-mRNA trans analysis by testing a normalization that corrects only technical covariates. Comparing these results showed that hardly any findings from the more conservative correlation are not found in the PEER-corrected analysis that also finds many additional associations. This indicates that even in this trans-analysis, PEER removes technical variation efficiently thus increasing overall power without correcting away biological effects.

Changes in the manuscript: We now discuss the effect of normalization on the results in Supplementary Methods section 8.

**Page 3: "challenge has been to analyze cellular phenotypes, such as gene expression, resulting in large catalogs of regulatory variants, known to affect many human diseases and traits"**
**It would be better to use one concrete example to show the power of genomic analysis to reveal disease causality.**

RE: We prefer to keep the introduction without examples in order to keep the manuscript within reasonable length but we are happy to add examples if considered necessary.

**Page 5: "As expected, the vast majority of the total transcription variation is among individuals within populations, with only 3% explained by population differences. Yet, between population pairs, we detect 263-4379 genes with significant differences in expression levels and/or transcript ratios"**
**First, please clarify if the 263-4379 genes constitute the 3% population difference? And how about the likelihood that the 3% across population differences were due to some statistical randomness?**

RE: The analysis of total transcription variation (the 3%) and the analysis of differentially expressed/transcribed genes are two separate methods to analyze the same question, the only thing in common being the same transcript quantifications. The 3% is an overall genome-wide estimate, and additional analysis permuting the population labels of the samples has confirmed that this is indeed far beyond what would be observed by chance.

Changes in the manuscript: We have revised the writing on page 5, and added a p-value ($<2.2 \times 10^{-16}$) for the population difference estimate.

**Page 7: "Even though these quantitative traits are expected to be biologically orthogonal, we find a significant enrichment of genes with both types of QTL (279 genes = 45% of trQTL genes = 2.15× enrichment, χ2 p < 2.2 × 10-16)."**
**Mechanisms of these two types of QTL are different, but this doesn't necessarily say if a gene is a QTL, it has to be subject to only one type of regulation. I would rather expect if a gene is a QTL, there might be a higher chance the gene is involved in more than one mechanism of regulation.**

RE: This is exactly the analysis that we describe in this section. We have tried to write this in a clearer manner.

Changes in the manuscript: We have revised the writing on page 7.

**Page 9: "In 13% of the genes the strongest eQTL variant is an indel, which is 37% more than for the matched null variants (Fisher exact test p = 5.6 Å~ 10-6; Fig S21), suggesting that indels are more likely to have functional effects than SNPs"**
**How many of this type of eQTL are protein-coding genes?**

RE: All the eQTL analysis is done on only protein-coding genes, as indicated on page 52 of the Supplementary Material, but this should indeed be specified in the main text as well.

Changes in the manuscript: We now mention on page 5 that eQTL analysis was done only on protein-coding genes.

**Section on "Characterization of regulatory variants"**
**In relation to how the authors connect the eQTL to functional sites I would urge them to look at the DNAse data (http://www.ncbi.nlm.nih.gov/pubmed/22307276). There have been many recent papers on this and I think that not discussing this data is a major limitation of the paper.**

RE: DNAse hypersensitive sites are included in our annotations, and the strong enrichment in them is described in Fig 2a, S18, S19, and we also refer to the Degner et al. paper in the text on page 9. We have added an analysis of the overlap of our eQTLs with their dsQTLs – showing a stong enrichment over the null – but we feel that a full replication of their excellent analysis is outside the scope of this paper.

Changes in the manuscript: We have added analysis of dsQTL overlap (pp 10 and Supplementary Methods section 11.5).