Author's response to reviews

Title: A longitudinal study of gene expression in healthy individuals

Authors:

Chris Karlovich (chris.karlovich@roche.com)
Guillemette Duchateau-Nguyen (guillemette.duchateau-nguyen@roche.com)
Andrea Johnson (andrea.johnson@roche.com)
Patricia McLoughlin (patricia.mcloughlin@roche.com)
Mercy Navarro (mercidita.navarro@roche.com)
Carole Fleurbaey (carole.fleurbaey@roche.com)
Lori Steiner (lori.steiner@roche.com)
Michel Tessier (michel.tessier@roche.com)
Tracy Nguyen (tracy.nguyen@roche.com)
Monika Wilhelm-Seiler (monika.wilhem-seiler@roche.com)
John Caulfield (john.caulfield@roche.com)

Version: 3 Date: 3 March 2009

Author's response to reviews: see over
March 2, 2009

Scott Edmunds Ph.D.
Senior Editor
BMC Medical Genomics

Dear Dr. Edmunds:

I am writing you regarding the reviews that we received from three referees for our manuscript entitled “A longitudinal study of gene expression in healthy individuals” (manuscript #1240624807190224). Please find in the following pages of this cover letter our point-by-point response to the referees’ comments. Additionally, I have uploaded a revised version of our manuscript onto your website. We have made several revisions in response to the referees’ comments that have added detail and clarity to the manuscript. We feel the current version adequately addresses their concerns and hope you will agree.

One concern raised in the reviews was whether informed consent was obtained from subjects enrolled in our study. The subjects were not part of a clinical trial where treatment was administered; they were enrolled in an observational study sponsored by Roche. I would like to assure you that an independent ethics committee approved the study protocol and that subjects gave written informed consent to participate in the study. The manuscript has been revised to clarify this. Additionally, concerns were raised about whether microarray data was deposited into one of the microarray databases. We are in the process of uploading the microarray data into the Gene Expression Omnibus (GEO) database. This is a high priority for us, and we will notify you as soon as we receive an accession number from GEO.

We thank you for your continued interest in our work. We look forward to hearing from you soon regarding a decision about the publication of our manuscript. Please let me know if you have any further concerns or questions.

Sincerely,

Chris Karlovich Ph.D.
Principal Scientist II
Roche Molecular Systems
chris.karlovich@roche.com
Reviewer’s report

Title: A longitudinal study of gene expression in healthy individuals

Version: 2 Date: 23 January 2009

Reviewer: Toni Whistler

Reviewer’s report:

This manuscript uses both microarray gene expression analysis and quantitative real-time RT-PCR to examine the variation of 11 immunomodulatory genes in healthy individuals over a 6 month period.

- Discretionary Revisions (which are recommendations for improvement but which the author can choose to ignore)

1. The paper mentions 80 subjects enrolled, of which 20 were taken for microarray studies and a separate 28 for the qPCR (8 of which are overlapped with the arrays), 18 had infections excluding them, 2 with adverse events. What of the other subjects?

   We initially examined gene expression by qPCR in all 80 subjects at each of the five time points for all analytes described in the study. However, an analysis of the data revealed that mean gene expression in the full group of 80 varied significantly by time point. (Please see (1) under major compulsory revisions for more details). We therefore repeated the study with a randomly chosen subset of 28 of the 80 patients where sufficient RNA remained.

   We present hematology data from all 80 subjects.

2. Results: When giving mean ± SD for the ages, both should be to the same number of decimal places, i.e. 34.5 ± 7.8 (as opposed to the SD being to 2 places.).

   This has been changed in the manuscript.

3. I personally would like to have the ages (mean ± SD) included into Table 1. It is nice to have all the demographics in one place. Makes for a quick and easy evaluation of the study design.

   We feel the purpose of Table I is to give the numbers of patients in each cohort. We did not intend to provide demographic information here. Specific demographic information about the subjects is given in the “Study population characteristics” section of the Results and Discussion.

4. Additional file 1: what do the dotted lines on the histograms represent?
The dotted lines represent the upper and lower clinical reference values for the individual cell counts displayed. A legend has been added to the Figure to indicate this.

-Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

1. Figure 1: Median gene expression – as measured by Ct? A label on the y-axis is needed. Also in the figure legend explanations are given regarding interpretation of the cat and whisker boxes, please extend this to include that the circles represent outliers.

   The y-axis label has been expanded so that it is more noticeable. The sentence following the box and whisker plot description does already note that outliers are indicated as additional points on the graph.

2. Figure 2: The title states “Immunomodulatory and reference gene expression over time”. In the graphs I see only 11 of the immunomodulatory genes.

   The figure legend has been edited to reflect the actual content of the figure.

3. Why is additional file 4 the first one mentioned in the manuscript – should this not be additional file 1?

   The manuscript has been changed so that Additional file 4 is now Additional file 1. The other Additional files have been renumbered to follow this one.

-Major Compulsory Revisions (which the author must respond to before a decision on publication can be reached)

1. RNA processing and quantitation: No mention is made of how the PAXgene tubes were batched for processing. Were tubes first stored (possibly frozen?), then RNA extracted at the same time? Were tubes processed on day of collection and then RNA stored? This detail is important for the reader to know. Was the batching of samples accounted for in the ANOVA? The first time this is addressed is the last paragraph of the paper, it needs to be addressed in the beginning so readers are aware of the shortcomings of the experimental design. Also no attempt has been made to show the batch effects on the data. I think that is important. There are ways of addressing technical issues in the analysis – batch effects can be accounted for as confounders in the ANOVA.

   We have amended the manuscript in the Methods section under “RNA processing and quantitation” to address the reviewer’s concerns about how the PAXgene tubes were collected and processed. We have
also amended the Methods section to address the reviewer’s questions about batch effects in the study design discussed above. Specific comments regarding these issues are restated below.

The PAXgene tubes were immediately frozen at -20 °C upon collection. The tubes were shipped frozen and stored at -80 °C for a period ranging from one week to six months before RNA was extracted. RNA extractions were performed in batches in the order in which the PAXgene tubes were received.

As mentioned previously, we did initially examine gene expression by qRT-PCR in all 80 subjects at each of the five time points for all analytes described in the study. An analysis of the data revealed that mean gene expression for all analytes varied significantly by time point in a way that was correlated across analytes. Differences in gene expression by time point can be attributed to either a true biological effect (e.g. a seasonal effect of gene expression) or an experimental bias (e.g. a “batch” effect resulting from processing the samples in batches as they were collected or running qRT-PCR assays on plates where all samples were from one or two time points).

Because we processed the samples and analyzed them by qRT-PCR in batches in the order in which they were received, we were unable to estimate batch effects in the data set independent of other biological and experimental variability. Further, since we could not estimate a batch effect, we could not properly construct a model that would adjust for it. Our solution was to repeat the entire qRT-PCR study as a 28-subject substudy. Samples collected at four time points in the substudy were entirely randomized. This design allowed us to (1) distinguish between biological and experimental variability, and (2) minimize experimental variability. No batch (i.e. plate) effect was needed in the mixed-effect modeling of the 28-subject data because the design was such that all observations being compared to each other were on a single plate (one plate per gene).

Note that we used a mixed-effect model, not ANOVA.

Data (CEL files + annotations) will be submitted to the GEO database. The accession number will be provided to the editors as soon as possible.
detailed. If a reader wanted to repeat this work they would need the modifications.

   The Methods section in manuscript has been updated to include details.

4. For the cRNA purification after the linear amplification the methodology states that either the Promega OR Qiagen kits were used. Why the mix in kit use? Was a comparison done comparing the impact of the different kits on the results?

   The reviewer’s concerns were addressed in the manuscript Methods section. A performance equivalency comparison was completed and is described.

5. Were labeling and hybridization controls included in the RNA sample preparation? The Affymetrix kits provide these controls. Were they (or something similar) used here? Was this part of the array QC? This is particularly important considering the low correlations that were found in your gene expression data.

   Control samples (Universal Human Reference RNA from Stratagene) were run with each processing run. Hybridization controls were used in the processing.

6. Quantitative RT-PCR: The RNA samples were quantitated using ribogreen for this assay, but by Bioanalyzer for the array methodology. Was this a sensitivity issue?

   We feel it is described clearly there that RNA quality for both qRT-PCR and array experiments was assessed on a Bioanalyzer instrument; RNA was quantified using the dye ribogreen.

7. What were the positive and negative controls run on each plate?

   One well of sample diluent served as a negative control for each sample run for the qRT-PCR experiments. Sample diluent consists of 10 mM TRIS pH 8.0, 0.1 mM EDTA, 0.020 mg/mL poly r(A) RNA (GE Healthcare), and 0.09% sodium azide w/v. “High input” and “low input” positive controls assessing expression of B2M were run on each plate. The target for the positive controls was human blood peripheral leukocyte total RNA (Clontech) (100 ng for high input and 1 ng for low input). Each plate was inspected to see that Ct values for high input and low input controls fell within a specified range.

   The Methods have been amended to include the above information.

8. Additional file 4. Primer data: It would be useful to the reader to add the expected size of the amplicon, and the primer efficiencies (or the assay
The expected sizes of the amplicons have been added to the table in Additional file 4. Assay efficiencies were determined to be between 93.0% (IL6R assay) and 105.9% (18S rRNA). This information has been added to the table legend.

9. What was the basis for the reference genes selected? “Typically used as reference genes” does not give enough information. Did the authors at least focus on genes generally used in blood? Choosing what is “typically” used is not a particularly strong scientific argument for the study design. The results of the RT-PCR are totally dependent on the housekeeping genes selected. They vary by tissue used. Therefore it is important to know a little more about why these were chosen.

We had 10 reference gene assays on hand in our laboratory and chose to evaluate the 4 described in the manuscript because these have been commonly evaluated in blood. The manuscript has been slightly revised to clarify this and two references (Silver et al, 2006, and Bas, et al ref 31) have been added to reflect that the reference genes chosen have been previously evaluated in blood.

10. Later in the paper: The authors discuss normalization by input RNA amount and how this appears to be a reasonable approach. However they do not discuss this in relation to cell numbers and possible shifts in the subsets of cells that their own data shows happens in whole blood. They have not convinced me that this would be better than normalizing to PPP1CA and B2M.

During the course of our analysis, we determined that all sources of variability, including any that might be due to heterogeneity of blood cell populations, was best minimized using the total RNA input. Adding a normalization using reference genes was not effective at further minimizing the experimental sources of variability. See comments with point 11 below for more details.

11. It is not clear to me how the reference genes were used to adjust the expression values of the immunomodulatory genes. This should be more clearly addressed in the methods. In the results they talk of normalizing against any of the reference genes individually, and that not making a difference to the variation. But which was chosen for the analysis? Today there are several analytical approaches that use 2-3 reference genes for the normalization (there are programs that will give you the best combination to use, and also normalize the data for you: geNORM, BestKeeper, qBase). Considering the reference gens were chosen because others had used them previously, not necessarily because they have been shown to be stably, constantly expressed in whole blood,
perhaps a more robust normalization should be considered.

The methods section has been amended to include a paragraph detailing the normalization analysis we performed and why we chose in the end to not draw conclusions from reference gene normalized data. We did initially normalize to a weighted average of four reference genes, which is generally thought to be a robust procedure. However, we concluded that this approach did not increase the precision in our data set and was therefore no better than normalizing only to total RNA input.

Why might this be? Our qRT-PCR assays used SYBR green to measure gene expression. While this is a common approach, SYBR green assays cannot be multiplexed, so normalization to one or more reference analytes cannot control for variation due to having the reference and target reactions take place in different wells. In addition, in our 28-subject design with one gene assayed per plate, the reference genes could not control for variability due to different plates either. Because the reference gene normalization did not reduce the overall variability in the data, normalizing to total input RNA only proved to be the best approach.

12. Analysis section: On page 9 the authors state “The significance of the effects was tested with t-tests, which necessitated a correction for the obtained P-values.” My understanding of array analysis is that the significance of effects has to be adjusted for the multiple testing, not because a parametric test was used. The next sentence correctly identifies FDR as a method used for correcting false positives. Please clarify what is meant.

We agree with the reviewer and have rewritten the sentence in the manuscript.

13. The Shapiro-Wilks test looks to see if the data is normally distributed. Only 17,239 probe sets were found to have a normal distribution. Seeing the ANOVA (linear mixed effect model) is a parametric test requiring normally distributed data. Why was the ANOVA applied to the 34,573 probe sets? Why was the data filtered to the 17,329 probe sets after the ANOVA?

Although we did not apply an ANOVA model, for the method that was used (a linear mixed effect model), it is necessary to check that model correctly fit the data. This was done by checking normality of residuals (observed – predicted signals) with Shapiro test, for example. Only probes that exhibit a normal distribution of residuals were then considered in the set of candidates (i.e. 17,329 probe sets) to be searched for differentially expressed genes. We did not use an ANOVA model because of the repeated measures structure in the data. The linear mixed effect model is a less restrictive model than ANOVA which assumes constant variance. The mixed effects model allows us to have different within and between subject variances which the ANOVA does not.
14. Further to this point, the authors’ state that the 17,329 probe sets were kept for further analysis. But in the very next sentence they performed correlations on the 34,573 probe sets. What was the further analysis performed on the 17,329 probe sets? Number corrected? What correlation coefficient was used? A Pearson (parametric) or a Spearman (non-parametric)

We used an intraclass correlation coefficient which is the ratio of the individual variance to the total variance. This coefficient is more appropriate than a Spearman correlation coefficient to assess the degree of association of the longitudinal data within subjects. To compute this coefficient we used the variances estimates computed with the linear mixed effect model.

15. Table 2: The hematology values for all subjects at all time points, is this the full 80 subjects in the study? Or just the 48 healthy subjects used to generate the qPCR and array data? Seeing the range of the WBC count is above normal I am assuming it includes those with the infections? For me the most relevant data would be for the 48 healthy subjects, as this is the major focus of the paper.

The values represent the values from all 80 subjects at all scheduled assessment times. Note that the infections usually involve only 1 or 2 time points for each patient and the numbers are too small to influence the means derived from 400 values. Further, all 80 subjects were by definition healthy at study entry although only 42 subjects were selected for the microarray and RT-PCR analysis.

16. Medication use is discussed in the results. It would be interesting to know if the paracetemol had any effect on the expression profiles. This is a much overlooked aspect of gene expression data and when one considers the number of clinical samples that are now being used in expression experiments an important one to address. If the information is available as to which time points the subjects were currently taking the medication, this would be a great analysis to include.

Paracetamol is indeed interesting. However, it is a generic drug with different strengths and formulations and is usually taken intermittently on no fixed schedule. The Cmax occurs about 90 minutes after dosing and the half life is about 2-3 hours. To study its effects on gene expression a trial in which the dosing, formulation, and time were controlled would have to be done. This was beyond the scope of the present study.

17. Hematology: The authors’ state “The mean cell counts were shifted to the left end of the normal laboratory range for each cell type.” However no explanation or postulation as to why. Any ideas? The data is available to compare the within person variability for those who remained healthy throughout the study, compared to those who were ill. This would add some perspective as to what
these numbers mean.

The term “shifted to the left” is inaccurate since the cell counts did not shift from a symmetrical to a skewed distribution. We have reworded the sentence to state, “Most subjects had cell counts that were in the lower half of the normal laboratory range for each cell type.” This observation is typically seen in clinical trials in healthy subjects and is probably the result of the upper limit of normal being set high so that patients who exceed it are truly ill, i.e., in most cases infected. As stated above, there are only one or two values that are abnormal for most patients with infection and this would not change the overall results shown in the table.

18. For the array data, differences between d90 and d180 showed 248 probe sets to be differentially expressed – was there enrichment in this list of genes for any particular biological process of molecular function? Looking at the biology may help in interpreting why there are these differences – it would hopefully come some indication as to whether the reason was seasonal affect, true biological effect or experimental bias. What were the levels of differential expression?

The manuscript has been edited to reflect more detail relating the gene list to biology. Specifically, the 248 probe sets correspond to 157 genes. Of those, 66 were associated with apoptotic activities. There was no specific enrichment of pathways but we found 66 genes having some apoptotic-related activities in publications. The log2-fold changes in these probe sets range from -0.9 to 0.6.

19. It is not clear to me what the authors really are getting at when they say “For each mRNA measured by qRT-PCR, gene expression in any given individual was less than 50% correlated across time points.” Please expand, how did you measure the correlation? What was the number/cutoff for what was correlated? Fewer than 50% were correlated across time points – 50% of what – individual genes?

This section of the manuscript has been modified to clarify this point. See changes in “within-individual variability in gene expression, qRT-PCR” section for details.

20. How do the authors account for the poor correlation between their gene expression and qPCR data?

The reviewer suggests that there is a poor correlation between the “gene expression” (we assume she is referring to microarray here) and qPCR data. The 28 subject qPCR and 22 subject microarray cohorts overlapped by only 8 subjects. We did not perform pairwise correlations between microarray and qPCR data on these 8 subjects because the
number was so small.

21. What is the % of probe sets showing a correlation > 4 if you only consider highly expressed genes? Say with signal intensities of log2 >10 or >14? We assume here that the reviewer is asking about probe sets showing a correlation > “0.4” not “4”.

   We observed a trend where genes expressed at the lowest levels generally showed poorer correlations than genes expressed at higher levels. We did not look at correlations specifically in the most highly expressed subset of genes. However, we found that for most of the highly expressed genes with a low correlation coefficient (<0.1) the corresponding probe sets were not of good quality (i.e. they did not map to the current version of the human genome sequence). (see figure 5).

22. Additional File 3: “Fold changes were calculated as 2 raised to the power of the estimated effect for gender or age group of that gene.” It is not clear to me how the fold change was calculated – perhaps an equation would clarify this?

   The caption has been changed to use mathematical notation to explain the fold change, as suggested by the editor. See manuscript for details.

23. It is not surprising that you are not showing any age-related changes – perhaps if you used slightly greater differences between your age groups. Your “younger group” includes people to 55 years, your “older” group >55. What if you use a group 18-35, and the >55?

   This is unlikely to show a difference. First about half of the lower age group would be excluded from the analysis resulting in increased variability. Second, the biological differences are potentially strongest at the cut-point chosen. After 55 men and women undergo large physical changes including bone loss, more rapid aging of the skin, loss of muscle mass, etc.

24. I personally find inclusion of the adverse event subjects, irrelevant. In my mind it should not be included in the study. The infection data is of interest, but I would like to see it put more in to the context of the variation that it may cause. Comparing it to what is seen in the subjects healthy through the study.

   We agree that the infection data has limited value because of the variability in type of infection and in sampling times. However, infections will occur in clinical trials and it is useful to have some data to support the contention that the overall effects of typical low grade infections on gene
expression will be limited.

**Level of interest:** An article of limited interest

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**
I declare that I have no competing interests
Reviewer’s report

Title: A longitudinal study of gene expression in healthy individuals

Version: 2 Date: 15 July 2008
Reviewer: Halima Moncrieffe

Reviewer’s report:

This study provides gene expression data from peripheral blood samples of healthy adult volunteers stratified by age and gender. This provides a useful data resource comparing qPCR and Affymetrix HG U133 Plus 2.0 microarrays. Blood was sampled at 5 time points, taken in the morning, post-fasting over a 6 month period. 28 individuals were sampled for Affymetrix arrays and a number of inflammatory genes were selected for qPCR.

MINOR ESSENTIAL REVISIONS:

1) P.5, If there is data for ethnicity of volunteers, this should be provided.

    Ethnicity data is now included in “Study Population Characteristics” within the Results and Discussion in the following sentence: "The healthy volunteers, 79 Caucasians and 1 Oriental, were from eastern France and stratified into four cohorts of 20 individuals each based on age (22 - 55 and >55) and gender (Table 1)."

2) Given that this is a longitudinal study, samples will have necessarily had differing times in storage. The authors should clarify how much time the sample spent as RNA and where it was stored until it was processed for microarray. E.g. were all samples stored in the Paxgene kit and then extracted just prior to processing for microarray?

    As described in Major Compulsory Revisions (1) (Reviewer TW), the RNA was extracted in several batches within six months of when it was received. Microarray processing was carried out in two batches as stated in the manuscript.

3) P11 – how were the subset of 28 individuals chosen?

    The 28 subjects were chosen at random from the larger set of 80 where sufficient RNA was available. This issue is discussed in detail in Major Compulsory Revisions (1) (Reviewer TW) and clarified in the Materials and Methods section of the manuscript.

4) Figure 1: Y-axis is obscured.
Y-axis label has been extended for readability.

5) Figure 3: 2 probe sets have fold change less than 0.5 and have a p value above the nominal Type I error cutoff. The text should clarify that this figure is prior to multiple testing correction.

A explanation has been added

6) P14: Due to microarray processing in two batches, it was not possible to compare the gene signal intensities over 6 months as perhaps implied in the abstract text, however the microarray data was compared for d90 to d180 i.e. 3 months and upto day 28 ie 1 month.

We have added more detail to the “Microarray Procedures” section of the Methods for clarity. Specifically, the samples were processed in two batches. The first batch included samples up to Day 28. The second included samples at Day 90 and Day 180. Therefore, we could compare Day 90 with Day 180 and Days 0, 14 and 28 with one another.

7) Figure 7: Y-axis is obscured.

Y-axis label has been extended for readability.

8) Additional file 1 needs a figure legend and to crop off the small print at the bottom of each histogram.

The figures have been replaced with a new set that do not contain the small print under each histogram. A legend has been added to the Figure.

DISCRETIONARY REVISIONS

1) p6. RIN < 6.5 is a low threshold. What was the average and range of RIN for those samples which were included in the study?

We did not determine the average RIN value for samples in the study, but we determined that RNA from 6.3% of samples was below our quality threshold of RIN = 6.5. These specimens were not included in the study.

We performed experiments on PAXgene RNA with RIN < 7 to determine a threshold where quantification of analytes was affected by RNA quality. In our tests, quantification of an analyte was significantly affected at RIN < 5. However, at RIN > 6, quantification did not seem to be affected. We therefore set our threshold at 6.5.

2) Why was an UNG digest necessary for this qPCR reaction.

Uracil-N-Glycosylase (UNG) was added to the reaction mixes to prevent carryover contamination (contamination of a PCR by the amplicon generated in a previous PCR). Our reaction mixes use dUTP instead of
dTTP. UNG cleaves dUTP. Therefore, if a PCR amplicon generated in one PCR contaminates another, UNG in the reaction mix will destroy the contaminating amplicon.

The manuscript has been amended in the Methods section to clarify the role of UNG.

3) P7. 2 ng of RNA, easier to follow than two ng of RNA

The manuscript has been amended to reflect this change.

4) P8 Reference for SAS Proc Mixed.

Has been added to manuscript.

5) P9 Quote total number of microarrays used, if present in at least one sample ie 1 of 112?

The manuscript has been revised to include the requested information.

6) P12 – do dotted lines refer to 95% Confidence intervals on additional file 1. Figure legends should be provided for the additional files.

See reponse to Dr Whistler, question number 4.

7) To clarify conclusion point 2, p19. Do the authors have data showing RNA extraction and storage using different Pax gene kit lot numbers and then extraction for gene expression profiling. I suspect the authors mean the batch effect on Affymetrix processing (yes) rather than RNA extraction, but this is not clear.

We have internal data from about 15000 PAXgene purifications with different kits, and we have control samples run at different times using different lots. The results are comparable between different kits and lots, and more importantly fall well within specs suggested by Qiagen)

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:

I declare that I have no competing interests
Reviewer’s report

Title: A longitudinal study of gene expression in healthy individuals

Version: 2 Date: 29 September 2008

Reviewer: martin petrek

Reviewer’s report:

Karlovich et al describe here the gene expression profiles in healthy men and women in whole blood and their variability over period of 6 months. They used two approaches (qRT-PCR and microarray) to assess the differences in expression profiles in subgroups according gender, age. This is an original article which addressed an important question in the field of gene expression, but there are several limitations with this study, which are addressed below.

Major compulsory revisions

1) The authors combined in this paper the Result and discussion. According opinion of this reviewer, in fact there is no discussion on the presented data. The authors acknowledge David Trollinger of Source Precision Medicine for discussions regarding the genes chosen for qRT-PCR, however they do not explain their choice to the readers. I also miss the explanation about the selection of cut-off for age. There is no discussion about the data and their practical importance. Their conclusions are not based on the data obtained and are to vague, or let me say general (e.g. Page 19-„These parameters can be used to estimate the number of subjects needed to observe significant differences from normal gene expression in clinical studies.“)

The genes chosen are based on known relevance to rheumatoid arthritis and autoimmune disease to benchmark a companion study looking at biomarker responses after initiating new therapies in patients with rheumatoid arthritis. The age cut-point was selected so that all women in the perimenopausal period would be in a single cohort. Since menopause can begin in the early 40s it was elected to include them in the younger cohort. Finally we believe that the data are important in helping to design clinical trials using gene expression as a pharmacodynamic readout.

2) It seems that no ethics committee permitted this study; the subjects did not sign the informed content. This is a VERY MAJOR FLAW!

All subjects were consented according to the Declaration of Helsinki and the protocol was reviewed and approved by the ethics committee and registered with the French health authorities. The following paragraph has been added to the Methods section: “The study was conducted in accordance with good clinical practice, the Declaration of Helsinki, and
appropriate regulatory guidelines. An independent ethics committee approved the protocol. Subjects gave written informed consent to participate in the study.”

3) The authors performed the microarray analysis and qRT-PCR separately; they did not combine the results. It seems that two different studies were performed and just joined together.

To address previous reviewer’s comments, we have amended the manuscript to describe in detail why and how the 28 subjects were chosen for the qRT-PCR study.

4) The study is presented as an analysis of 80 subjects (they called this “main” study), however, one ¼ (25%) were analysed in fact. Especially for qRT-PCR is the number very low-qRT-PCR was performed in only 22 subjects (subgroups: 8 young females, 7 older females, 5 young males, 8 older males). There is a very low overlap of samples used for both techniques (“Samples from eight subjects were analyzed on both qRT-PCR and microarray”). Especially, when they used such little RNA per qRT-PCR reaction (10 ng)? We recommend to analyse the expression profile by qRT-PCR in all enrolled subjects.

We did not have enough sample material from all subjects to redo the 80-subject study. We therefore chose to repeat the qPCR work on a random subset of 28 individuals from the group of 80 where sufficient RNA still remained for qPCR. We chose a 28-subject subset because this was the maximum number of subjects we could profile at four time points in triplicate on a single plate. Use of a single plate eliminated plate-to-plate variability.

5) RNA quality was assessed on an Agilent Bioanalyzer 2100. Why the authors did not quantify the RNA samples by this technique, they wrote that “samples were quantified using ribogreen, a nucleic acid stain“.

At the time of this study, quantification of RNA using a Bioanalyzer 2100 was shown by us to have less precision and accuracy than quantification based on a standard curve using ribogreen or by UV spectrophotometry on a Nanodrop instrument.

More recently, we have determined that newer versions of the Agilent Bioanalyzer software quantify RNA with much better precision and accuracy than in the past. However, the precision of measurements is still not quite as good as on a Nanodrop instrument.

6) Expertise from statistician is needed.

Three statisticians, Duchateau-Nguyen, Johnson, and Navarro are authors of the manuscript.
7) All qRT-PCR data are presented as raw Ct. However, the intra-and inter-assay variability may have influence on the presented data. Did the authors use any calibrator?

A calibrator is useful in helping to correct for variability introduced by plate-to-plate differences. However, all data for a given analyte in the 28 subject qRT-PCR study was generated on a single plate. With our normalization approach, there were no plate-to-plate comparisons to reference gene analytes on other plates. Therefore, while each plate included a calibrator, we did not need to use it for normalization.

How was the amplification efficiency in all genes?

As mentioned previously, our amplification efficiencies ranged from 93.0% (IL6R assay) to 105.9% (18S rRNA). This information has been added to the table legend for Additional File 1.

I recommend presenting the data as ratio to reference gene, not as Ct raw data.

As we discussed previously, the variability in the qRT-PCR data set was not reduced when reference genes were used compared with total RNA input as a normalizer.

As described previously and in the manuscript, our qRT-PCR assays are SYBR Green assays. SYBR Green assays cannot be multiplexed, so the target and reference assays are assembled as separate reactions in different wells of a plate or even on different plates. Given this, it is not so surprising to us that normalizing by total RNA input introduced less variability than normalizing by any of the candidate reference genes.

How significant was the variation between several lots of chemicals (page 19-“A measurable bias was introduced when samples from the present study were processed and analyzed with different reagent lots over time in sequential batches (data not shown). The authors mentioned this without showing the data (please make additional file from this).

For the qRT-PCR work, a “batch effect” bias was introduced as already described in the original 80-subject study. However, the 28-subject study described in the manuscript was performed with one lot of reagents.

For the microarray study, the extent of technical bias (i.e. batch and reagent effects) could not be measured in this study as a time effect and technical bias are confounded. We therefore have removed the word “measurable” from the sentence “A measurable bias was introduced....”

Minor essential revisions
1) I would recommend to change the title to” ...healthy non-smokers”. In the
subject description it should be defined more precisely if there were some ex-smokers enrolled into this study.

We did not capture the former smoking status of subjects enrolled in this study.

2) Page 1-Abstract: Results about the patients with cancer and anemia should not be main-results from the study. Two patients are not enough for conclusions.

We respectfully disagree since the gene expression changes were so strong and they were genes that made “biological sense”. They also agree with other reports.

3) Please give more information about how much RNA was obtained, how many ml of blood were taken to PAXgene tube.

The mean yield of RNA from all extractions that passed the quality control step (RIN ≥ 6.5) was 7.0 μg per tube (1 SD = 2.6 μg). 2.5 mL of blood are drawn into a PAXgene tube which contains 6.9 mL of additives. This information has been added to the Methods section of the manuscript.

4) Page 12-„shifted to the left end of the normal laboratory range“. Please define the left end.

Please see response to Dr Whistler.

5) Page 7. „……with minor modifications”. However, the used modification is not mentioned.

Modifications have been made to the manuscript to address the reviewer’s concerns.

6) Selected genes were measured by qRT-PCR at four time points (Baseline, Day 28, Day 90 and Day 180), however they obtained samples from „five scheduled visits on Days 1, 14, 28, 90, and 180 per protocol.“

As mentioned previously, we chose to measure gene expression by qRT-PCR at what we felt were the four most informative time points in the 28-subject study. Use of four time points instead of five allowed us to profile all 112 subject time points in triplicate plus controls on a single plate.

7) Additional file 3-The data are given in Fold Change (which group was set to 1?)

The caption has been modified to clarify this point.

8) Additional file 2-Norm = normalized values, it is not mention against which gene is normalised

Four reference genes were used. The caption has been modified
to clarify this.

9) Table 2 -Hematology values of all subjects over all time points—it would be more interested to see changes in individuals over six months

We have examined the data graphically. Within-subject variability is low as shown quantitatively in Table 2.

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Needs some language corrections before being published

Statistical review: Yes, but I do not feel adequately qualified to assess the statistics.

Declaration of competing interests:
I declare that I have no competing interests. MP