Author's response to reviews

Title: A pooling-based genome-wide analysis identifies new potential candidate genes for atopy in the European Community Respiratory Health Survey (ECRHS)

Authors:

Francesc Castro-Giner (fcastro@creal.cat)
Mariona Bustamante (mariona.bustamante@crg.es)
Juan Ramon González (jrgonzalez@creal.cat)
Manolis Kogevinas (kogevinas@creal.cat)
Deborah Jarvis (d.jarvis@imperial.ac.uk)
Joachim Heinrich (joachim.heinrich@helmholtz-muenchen.de)
Josep-Maria Antó (jmanto@creal.cat)
Matthias Wjst (m@wjst.de)
Xavier Estivill (xavier.estivill@crg.es)
Rafael de Cid (decid@cng.fr)

Version: 2 Date: 30 July 2009

Author's response to reviews: see over
Catherine Laprise

Major comments

Comment 1: My first comment concerns the pooling itself. It is probable a good alternative to pool samples to reduce dugbet expenses. However, I think it’s necessary to perform analyses for some groups of samples in order to have an idea of biological variation (for example: analyze three pools of 25 samples per phenotype). It is also important to duplicate some analyses to verify the technical variability.

Response:
We agree with the reviewer that biological variability could have a role during the pooling strategy. The idea of performing different ‘subpools’ according to phenotypes is interesting and could capture more variability. However, the issue of false positives (that is probably the main concern in this analysis) is not relevant to the issue of having large pools, since this is controlled in the next phase by doing individual analyses. We acknowledge that with having a unique and larger pool for each phenotype we may be losing positive signals i.e. have false negatives but we consider this as a less crucial issue for this analysis. Pooling stage was designed to obtain experimental candidate regions, and despite loosing true signals, major factors should be identified from such strategy. Some of the results obtained from this analysis validate the value of this strategy. We have added a sentence on this issue in the discussion regarding technical variability, we have performed several quality control approaches to eliminate possible technical bias:

- Eight individuals from our databank and one HapMap subject with known genotypes were genotyped with the same Illumina chip and procedure. SNPs incoherent with the previous genotype data were excluded (n= 296, 0.09%). The percentage of excluded was similar to other studies [Abraham 2008].
- One pool of 57 subjects from HapMap was used to validate the method of allele frequency estimation. Those SNPs with a difference between real and estimated allele frequency in the pools >12% were excluded from the analyses (n= 3271, 1.03%). This value was selected according to previous reports [Steer 2007 and Craig 2005].
- Three independent replicates for each pool were constructed in order to account for errors in pool construction and genotyping. This error (sampling error) was accounted as the variance of the frequency between pools [Visscher 2003]. The formula for measuring the difference of frequencies between cases and controls accounts for this error [Visscher 2003].

We acknowledge that some of the points mentioned above concerning technical variability were not clear in the methods section and we have expanded this part in the revised manuscript (page 10).

References

Comment 2: As the authors note in their discussion, the samples does not reach significant levels to make sure that the associations was not a “false positive” one. The authors also note the second major weakness of their design concerning the lack of saturation of the region in term of number of genotyped SNPs. Indeed, I think that the authors need to increase the number of subjects in their association study (or perform a validation of significant association in an independent cohort) as well as the number of SNPs to dissect haplotype.

Response to comment on saturation
The comment about the lack of saturation in the discussion does not refer to the number of SNPs to saturate the regions selected, but for the regions that were not covered the second stage (chr5 and chr1, see Table 2). These regions contain segmental duplications and putative insertions or deletions [Kidd 2008]. They were not included in the second stage because their complexity could be a source of bias due to non-classical inheritance of markers located on them. To avoid misunderstanding regarding the genotyping coverage, the concept of saturation has been clarified in the manuscript.

Response to comment on sample size
We agree that small sample size is a limitation of this kind of studies and we already commented on this in the discussion. However, we disagree in part with the comment by the reviewer on significant levels. We should note that, in the second stage, significance levels resist conservative Bonferroni corrections (at least for some markers in SGK493 and MAP3K5 SNPs) and this reduces the probability of false positives. These significance levels achieved indicate that this sample is powered to detect large effects. In the revised paper we have included formal Bonferroni corrections for the second stage, have included a supplementary table showing the power of the study for replication and have modified extensively the discussion of these issues in the manuscript as indicated here below.

For the genome-wide analysis on pooling-based DNA, a strict correction for multiple testing was performed. We used false discovery rate (FDR) at 5%, which produced a very low p-value threshold of significance (< 8x10^{-7}) similar to other GWAS analysis on individual samples. However, given the small number of subjects and the corrections applied, we were powered to detect larger effects and under-powered to detect other signals (false negatives).
Concerning the second phase of the analysis, we stated in the discussion that one of the limitations is the relatively small sample size for replication and multiple testing corrections were initially not performed. In the revised manuscript we did use Bonferroni correction and still identified statistically significant results. We also performed an estimation of the statistical power for replication in the additional sample of 429 cases of atopy and 222 controls given the parameters (odds ratios and allele frequencies) obtained in the individual analysis of the pooling samples (see the below Table). For the associations listed in Table 3 of the manuscript, we are completely powered to detect associations in the replication sample. Power was calculated a priori (using 429/222) and a posteriori, after data curing. As indicated above we included a sentence in the manuscript and a table in the supplementary material.

Table. Statistical power calculation for replication using Quanto software v1.2.4 (http://hydra.usc.edu/gxe) for significance (two-sided) of 0.05 and additive genetic model.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Gene</th>
<th>MAF</th>
<th>Odds Ratio</th>
<th>Case/control</th>
<th>Power</th>
<th>A priori power</th>
<th>Effective power</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs11124858*</td>
<td>SGK493</td>
<td>0.367</td>
<td>0.43</td>
<td>429/222</td>
<td>0.99</td>
<td>331/178</td>
<td>0.99</td>
</tr>
<tr>
<td>rs13409978*</td>
<td>SGK493</td>
<td>0.122</td>
<td>0.22</td>
<td>429/222</td>
<td>0.99</td>
<td>334/178</td>
<td>0.99</td>
</tr>
<tr>
<td>rs4952590**</td>
<td>SGK493</td>
<td>0.141</td>
<td>0.18</td>
<td>429/222</td>
<td>0.99</td>
<td>317/163</td>
<td>0.99</td>
</tr>
<tr>
<td>rs1440095</td>
<td>SGK493</td>
<td>0.379</td>
<td>0.50</td>
<td>429/222</td>
<td>0.99</td>
<td>333/172</td>
<td>0.99</td>
</tr>
<tr>
<td>rs10934938</td>
<td>COL29A1</td>
<td>0.23</td>
<td>0.57</td>
<td>429/222</td>
<td>0.96</td>
<td>296/150</td>
<td>0.87</td>
</tr>
<tr>
<td>rs9402839*</td>
<td>MAP3K5</td>
<td>0.138</td>
<td>2.39</td>
<td>429/222</td>
<td>0.99</td>
<td>314/166</td>
<td>0.99</td>
</tr>
<tr>
<td>rs9494554**</td>
<td>MAP3K5</td>
<td>0.095</td>
<td>2.47</td>
<td>429/222</td>
<td>0.99</td>
<td>329/175</td>
<td>0.99</td>
</tr>
<tr>
<td>rs12483377*</td>
<td>COL18A1</td>
<td>0.108</td>
<td>4.11</td>
<td>429/222</td>
<td>0.99</td>
<td>282/137</td>
<td>0.99</td>
</tr>
</tbody>
</table>

*SNPs found associated in the pooling-based GWA or their perfect tags (r² ≥ 0.8) used as substitutes in the genotyping design.
**SNPs that remained significant after a 5 % FDR in the pooling-based GWA

Finally, regarding the suggestion to increase the number of individuals in the replication sample or validate these results in an independent cohort, we completely agree with both reviewers that these options would increase the validity of results. However, we consider that results obtained in the pool are successfully replicated in the second phase. As we noted above, complementary sample was statistically powered to detect most of the associations reported in the first stage. In the discussion and conclusions we encourage the replication of these findings by other groups.

References
Florence Demenais

Comment: SNPs within MAP3K5, COL29A1 and COL18A1 genes were reported to be associated with atopy and/or atopic asthma, although the alter results were not always consistent between those obtained from DNA pooling and individual genotyping. Can the authors explain these inconsistencies?

Response:

We agree with the reviewer that we should have discussed more completely these results and have modified the discussion accordingly. A detailed response can be found here below:

We acknowledge that findings in MAP3K5, COL29A1 and COL18A1 genes are not completely replicated in the three steps of this paper. Given the potential biological plausibility of these genes to be associated with asthma or atopy, we consider useful for future research to report them in the manuscript, although remarking the inconsistencies observed.

Non-replication and inconsistency of initial findings is a common feature in the search of genetic determinants in candidate genes and GWA studies for complex phenotypes, even though strict p-value threshold were applied. Possible causes of non-replication are many [Ioannidis 2007], but we did not identify a specific cause for these inconsistencies. As noted by the reviewer in the next comment, a bias could be caused by the limited number of individuals included in the pool. However, the statistic used for measuring the significance of differences in allele frequencies, accounts for the number of subjects included in the pool construction. Other limitation inherent to the DNA pooling strategy is the impossibility of adjustment for potential confounders. Statistical analysis for individual data was more strict and was performed using logistic models adjusted by the most common confounders in asthma and atopy. So, differences between pooling and individual analysis may be driven by the effect of some of these confounders in the pool samples although differences in the descriptive analysis (Table 1) were not large. Regarding the replication stage, phenotypic heterogeneity is an unlikely source of bias given the detailed and homogeneous characteristics of the ECHRS protocols concerning the definition of asthma. The limited sample size in the replication sample could obscure minor effects, but the sample was large enough to identify major effects (see response to next comment).

References


Comment: We are concerned with the small sample size for a GWAS: 75 individuals in each of three groups examined (subjects with atopy, subjects with asthma and atopy, controls) using DNA pooling and the replication samples include at most 429 subjects in one group. It would be desirable to increase the sample size for this study or to use other samples for replication.
Response:

(Same comment as Reviewer 1)-

We agree that small sample size is a limitation of this kind of studies and we already commented on this in the discussion. However, we disagree in part with the comment by the reviewer on significant levels. We should note that, in the second stage, significance levels resist conservative Bonferroni corrections (at least for some markers in SGK493 and MAP3K5 SNPs) and this reduces the probability of false positives. These significance levels achieved indicate that this sample is powered to detect large effects. In the revised paper we have included formal Bonferroni corrections for the second stage, have included a supplementary table showing the power of the study for replication and have modified extensively the discussion of these issues in the manuscript as indicated here below.

For the genome-wide analysis on pooling-based DNA, a strict correction for multiple testing was performed. We used false discovery rate (FDR) at 5%, which produced a very low p-value threshold of significance ($< 8 \times 10^{-7}$) similar to other GWAS analysis on individual samples. However, given the small number of subjects and the corrections applied, we were powered to detect larger effects and under-powered to detect other signals (false negatives).

Concerning the second phase of the analysis, we stated in the discussion that one of the limitations is the relatively small sample size for replication and multiple testing corrections were initially not performed. In the revised manuscript we did use Bonferroni correction and still identified statistically significant results. We also performed an estimation of the statistical power for replication in the additional sample of 429 cases of atopy and 222 controls given the parameters (odds ratios and allele frequencies) obtained in the individual analysis of the pooling samples (see the below Table). For the associations listed in Table 3 of the manuscript, we are completely powered to detect associations in the replication sample. Power was calculated a priori (using 429/222) and a posteriori, after data curing. As indicated above we included a sentence in the manuscript and a table in the supplementary material.

Table. Statistical power calculation for replication using Quanto software v1.2.4 (http://hydra.usc.edu/gxe) for significance (two-sided) of 0.05 and additive genetic model.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Gene</th>
<th>MAF</th>
<th>Odds Ratio</th>
<th>Case/control</th>
<th>Power</th>
<th>Case/control</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs11124858*</td>
<td>SGK493</td>
<td>0.367</td>
<td>0.43</td>
<td>429/222</td>
<td>0.99</td>
<td>331/178</td>
<td>0.99</td>
</tr>
<tr>
<td>rs13409978*</td>
<td>SGK493</td>
<td>0.122</td>
<td>0.22</td>
<td>429/222</td>
<td>0.99</td>
<td>334/178</td>
<td>0.99</td>
</tr>
<tr>
<td>rs4952590**</td>
<td>SGK493</td>
<td>0.141</td>
<td>0.18</td>
<td>429/222</td>
<td>0.99</td>
<td>317/163</td>
<td>0.99</td>
</tr>
</tbody>
</table>
Finally, regarding the suggestion to increase the number of individuals in the replication sample or validate these results in an independent cohort, we completely agree with both reviewers that these options would increase the validity of results. However, we consider that results obtained in the pool are successfully replicated in the second phase. As we noted above, complementary sample was statistically powered to detect most of the associations reported in the first stage. In the discussion and conclusions we encourage the replication of these findings by other groups.

Comment: As pointed out by the authors in the discussion, there is no correction for multiple test in the analyses carried out at the individual level. Even without that correction, the p-values shown in Table 3 are not really small, except for SGK493 SNPs and one MAP3K5 SNP. This should be indicated in the results which could mainly focus on SKG493.

Response: (see also response to previous comment)
We agree with the reviewer. We did not perform multiple testing corrections in the analysis of individual data because we had previous evidence of association obtained in the pooling based approach. We consider that replication is essential for establish the validity of an association, more than multiple testing correction that usually can mask real positive results (false negative). Markers in SGK493 and MAP3K5 genes are significant even after a Bonferroni correction (p-value threshold 0.05/46 effective number of SNPs = 0.001) which is widely accepted as a conservative method. We included Bonferroni levels in the manuscript. P-values for the other markers do not remain significant after the Bonferroni correction, but since p-values for replication at the second stage are nominally significant, they are also presented.

We also appreciate the comment of the reviewer about focusing the results and discussion on the SGK493 gene. We have extended the analysis of SGK493 with expression data. The discussion was modified focusing the text on this gene.

Comment: Since Figure 1 is not easy to read, the pairwise LD measures (D’, r2) among SGK493 SNPs could be indicated in the text. Multiple regression with these SNPs might
be performed to determine which SNPs have independent effects. Haplotype analysis might be also carried out.

Response:
We agree with the reviewer. We modified Figure 1 with the information on LD only among the evaluated SNPs in this analysis (not HapMap). Only LD measures (D’ and r^2) between most significant SNPs from haplotype analysis were included in the text. LD information for the rest of SNPs combinations were not annotated in the text because we considered that would be difficult to read due to the high number pairwise measures. These data could be found in the new Figure 1. The previous figure showing LD patterns in HapMap subjects has been moved to supplementary material (supplementary Figure 1).

The haplotype analysis has been done and reported in the new version of the manuscript, although we consider that results did not modify the conclusions obtained from single marker analysis. We also performed the multiple regression analyses (table below). This analysis supports the evidences on the rs4952590 marker. However, we did not include this analysis in the main paper because the use of multiple regression analysis may not be appropriate in this case. The different loci of SGK493 are in strong LD. Multiple regression-like analyses take each variable as independent, which can produce misleading results with correlated data.

<table>
<thead>
<tr>
<th>SAMPLE</th>
<th>Best model SNPs variables</th>
<th>Estimate</th>
<th>Standard deviation</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pooling</td>
<td>rs4952590</td>
<td>-0.29</td>
<td>0.08</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>rs2424</td>
<td>-0.11</td>
<td>0.06</td>
<td>0.09</td>
</tr>
<tr>
<td>Replication</td>
<td>rs4952590</td>
<td>-0.38</td>
<td>0.13</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>rs13409978</td>
<td>0.27</td>
<td>0.13</td>
<td>0.04</td>
</tr>
<tr>
<td>Combined</td>
<td>rs4952590</td>
<td>-0.31</td>
<td>0.11</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>rs13409978</td>
<td>0.16</td>
<td>0.11</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Comment: The discussion regarding genes for which there is no strong evidence of association (eg COL18A1) should be shortened.

Response:
We agree and have done shortened the discussion regarding the role of COL18A1 and focused discussion on SGK493.