

1 **Copy number variation of scavenger-receptor cysteine rich domains within *DMBT1* and Crohn's**  
2 **disease.**

3 **Running title: *DMBT1* and Crohn's disease**

4 Shamik Polley (1), Natalie Prescott (2), Elaine Nimmo (3), Colin Veal (1), Ida Vind (4), Pia Munkholm  
5 (4), Peder Fode (4), John Mansfield (5), Paal Skyt Andersen (4), Jack Satsangi (3), Christopher  
6 Mathew (2), Edward J Hollox (1\*)

7 1. Department of Genetics, University of Leicester, Leicester, UK

8 2. Department of Medical and Molecular Genetics, King's College London, London, UK

9 3. Centre for Genomic and Experimental Medicine, University of Edinburgh, Edinburgh, UK

10 4. Microbiology and Infection Control Unit, Statens Serum Institut, Copenhagen, Denmark

11 5. Institute of Genetic Medicine, Newcastle University, Newcastle Upon Tyne, UK

12

13 **\* Corresponding author**

14 Dr Edward Hollox

15 Department of Genetics

16 Adrian Building

17 University of Leicester

18 University Rd

19 Leicester

20 LE1 7RH

21

22 **Keywords:**

23 Crohn's Disease, DNA Copy Number Variations, Case-Control Studies, Genomics, Agglutinins

24

25

26 **Abstract**

27 Previous work has shown that the gene *DMBT1*, which encodes a large secreted epithelial  
28 glycoprotein known as salivary agglutinin, gp340, hensin or muclin, is an innate immune defense  
29 protein that binds bacteria. A deletion variant of *DMBT1* has been previously associated with  
30 Crohn's disease, and a *DMBT1*<sup>-/-</sup> knockout mouse has increased levels of colitis induced by dextran  
31 sulphate. *DMBT1* has a complex copy number variable structure, with two, independent, rapidly  
32 mutating copy number variable regions, called CNV1 and CNV2. Because the copy number variable  
33 regions are predicted to affect the number of bacteria-binding domains, different alleles may alter  
34 host-microbe interactions in the gut. Our aim was to investigate the role of this complex variation in  
35 susceptibility to Crohn's disease by assessing the previously reported association. We analysed the  
36 association of both copy number variable regions with presence of Crohn's disease, and its severity,  
37 on three case-control cohorts. We also reanalysed array comparative genomic hybridisation data  
38 (aCGH) from a large case-control cohort study for both copy number variable regions. We found no  
39 association with a linear increase in copy number, nor when the CNV1 is regarded as presence or  
40 absence of a deletion allele. Taken together, we show that the *DMBT1* CNV does not affect  
41 susceptibility to Crohn's disease, at least in Northern Europeans.

## 42 Introduction

43 Crohn's disease (CD) is a chronic debilitating inflammatory disease that most frequently affects the  
 44 terminal ileum but can affect any part of the gastrointestinal tract. In the West, CD has a prevalence  
 45 of around 150 per 100,000<sup>1</sup>, with environmental and genetic variations making an approximately  
 46 equal contribution to disease risk<sup>2</sup>.

47 The most recent progress in elucidating the genetic variation responsible for CD has come from SNP-  
 48 based genomewide association studies, which have identified 163 loci which contribute to the  
 49 genetic risk of the disease<sup>3,4</sup>. Nevertheless, even with well-powered analyses of 15,000 cases and  
 50 15,000 controls, only 13.6% of disease variance has been explained, suggesting that other genetic  
 51 risk variants exist that are not interrogated by current SNP-GWAS approaches. Copy number  
 52 variation, where a whole or part of a gene differs in copy number in different individuals, is a  
 53 potential candidate type of variation that is often not well-tagged by flanking SNP alleles<sup>5</sup>. CNV of  
 54 the beta-defensin genes and amylase has been shown to affect susceptibility to psoriasis and obesity  
 55 respectively<sup>6,7</sup>, indicating that CNV can contribute to genetic variance of common complex diseases.  
 56 A genome-wide association study directly interrogating CNV by array CGH identified a CNV in the  
 57 HLA region and at the *IRGM* gene associated with CD<sup>8</sup>, but other more complex CNVs may also be  
 58 associated with CD.

59 Because methods to reliably and cost-effectively type CNVs genomewide are lacking, recent  
 60 literature has focused primarily on studies of candidate genes chosen for their known role in the  
 61 etiology of CD. One example is association of CNV of the beta-defensin gene region with CD, where  
 62 an initial study supported an association of low beta-defensin copy number with CD, an effect only  
 63 seen in colonic CD rather than ileal CD<sup>9</sup>. A subsequent study on a larger cohort of cases and controls  
 64 found a significant association but in the reverse direction<sup>10</sup>, and both studies are limited by low  
 65 statistical power and limitations in the technology used to type CNV<sup>11,12</sup>. Indeed, the large  
 66 genomewide arrayCGH association study of CD patients and controls found no evidence of  
 67 association with beta-defensin CNV<sup>8</sup>, a finding supported by a rigorous study which also showed  
 68 that real-time quantitative PCR methods often used to type CNV could easily generate false positive  
 69 associations of CNV and disease<sup>11</sup>. This emphasised the importance of robust copy number  
 70 detection methods that minimised the chance of false positive results.

71 Another CNV that has been associated with CD is a deletion of part of the *DMBT1* gene, called  
 72 *DMBT1*<sup>SR47-13</sup>. CD patients have been shown to have a higher frequency of this deletion compared to  
 73 controls. *DMBT1* is a particularly attractive candidate gene, as it encodes a large secreted

glycoprotein (also known as salivary agglutinin, gp340, hensin or muclin) that is expressed in the gastrointestinal tract and is upregulated in CD<sup>14</sup>. Furthermore, *DMBT1* binds a wide variety of Gram positive and Gram negative bacteria, at least in saliva and the lung<sup>15,16</sup>, and *DMBT1* knockout mice show subtly enhanced sensitivity to experimentally-induced colitis<sup>13</sup>, although this has not been confirmed with an alternative *DMBT1* knockout mouse<sup>17</sup>. Nevertheless, the evidence suggests that *DMBT1* has an important innate defence function in the gastrointestinal tract.

The canonical *DMBT1* protein is composed of a regular array of 13 scavenger receptor cysteine-rich (SRCR) domains interspersed with SID domains, and followed by a CUB domain, a diverged SRCR domain, a further CUB domain then finally a zona pellucida domain<sup>18</sup>. The polymorphic *DMBT1*<sup>SR47-</sup> deletion previously associated with CD leads to the loss of four SRCR domains (SRCR3-6)<sup>19,20</sup>, (Figure 1), and since these SRCR domains have been shown to contain the binding sites for bacteria<sup>15</sup>, it has been suggested that the deletion leads to a quantitative change in the ability of *DMBT1* to bind bacteria, limiting the protection of the host mucosa against intestinal flora and therefore contributing to the pathogenesis of CD<sup>13</sup>.

Recent work has demonstrated that the *DMBT1*<sup>SR47-</sup> polymorphic deletion is in fact part of a wide spectrum of alleles affecting the copy number of SRCR domains within *DMBT1*<sup>20</sup> (Figure 1). Specifically, at the locus where the polymorphic deletion occurs (termed CNV1) there is also a polymorphic duplication allele of the same 4-SRCR domain repeat unit. Furthermore at the C-terminal SRCR region there is a further CNV (termed CNV2) where a single SRCR domain unit can vary between 0 to 11 copies per diploid genome. Taken together, this indicates that although the canonical *DMBT1* 13-SRCR array structure represents a common genotype containing 26 tandemly-arrayed SRCR domains per diploid genome (2x 13 SRCR arrays), in reality a wide range of SRCR domain numbers have been observed within *DMBT1* ranging from 14 to 40 SRCR domains per diploid genome. Therefore, through allelic variation alone, *DMBT1* molecules have the potential to contain between 7 and 20 SRCR domains, as a conservative estimate<sup>20</sup>.

Given this extensive variation, and the robust methods used to type it on small amounts of genomic DNA, we endeavoured to firstly replicate the original observation of an association of SRCR copy number on Crohn's disease on three large Northern European case-control cohorts, and to extend the analysis to the full allelic spectrum of *DMBT1* SRCR domain variation which might explain a significant amount of the genetic variance in CD susceptibility.

## Methods

106 *Danish cohort*

107 DNA was extracted from peripheral blood of native Danish CD patients recruited from a well-defined  
 108 geographical region (Copenhagen capital area, Denmark) during a two-year period from January 1,  
 109 2003 to December 31, 2004. The details of the Danish CD cohort are described elsewhere<sup>21,22</sup>. DNAs  
 110 from healthy blood donors from the Danish national blood bank were included as controls.

111 *Scottish cohort*

112 DNA was isolated from peripheral blood from CD patients were collected at the Western General  
 113 Hospital, Edinburgh, Scotland, which is a tertiary referral centre for IBD in South-East Scotland.  
 114 Detailed description of the Scottish cohort is given elsewhere<sup>11</sup>. Written consent from CD patients  
 115 was obtained prior to inclusion in the study. DNA from blood samples from unrelated  
 116 spouses/friends of IBD patients or samples obtained from the Scottish Blood Transfusion Service  
 117 were used as healthy controls. The study protocol was approved by Medicine and Oncology  
 118 Subcommittee of the Lothian Local Research Ethics Committee (LREC 2000/4/192).

119 *English cohort*

120 White European patients with CD were recruited from specialist IBD clinics in London and Newcastle  
 121 as reported elsewhere<sup>23</sup> after informed consent and ethical review (REC 05/Q0502/127). Patients  
 122 were recruited from Guy's and St. Thomas' Hospitals London, United Kingdom, St. Mark's Hospital  
 123 London, United Kingdom, and the Royal Victoria Infirmary, Newcastle, United Kingdom after ethical  
 124 review and informed consent from CD patients. Human random control (HRC) DNA samples from  
 125 lymphoblastoid cell lines derived from UK Individuals (from the ECACC collection held by Public  
 126 Health England: <http://www.phe-culturecollections.org.uk/>) were used as control samples.

127 *Copy number typing*

128 We used our extensively validated and robust paralogue ratio test approach to type copy number on  
 129 genomic DNA samples, as described previously<sup>24 20</sup>. Briefly, test and reference amplicons are  
 130 generated using the same PCR primer pair, one primer fluorescently labelled. The primers are  
 131 designed so that test and reference amplicons can be distinguished by a small difference in product  
 132 length by capillary electrophoresis on an ABI 3100xl. Eight positive control DNAs from the HapMap  
 133 panels were run on every plate to act as calibrators (supplementary table 1). Our previous study,  
 134 using repeat testing of identical samples, estimates the experimental error rate of CNV1  
 135 determination to be 0.37% and of CNV2 to be 0.33%<sup>20</sup>. WTCCC data was provided courtesy of the

WTCCC Access Committee and Dr Matthew Hurles (Wellcome Trust Sanger Institute). Raw data has been deposited with dbVar <http://www.ncbi.nlm.nih.gov/dbvar> accession number nstd77.

### *Statistical analysis*

Raw copy number data from PRT was normalised to have a standard deviation of one across the cohort. Data from cases and controls were analysed together, and, following visual inspection of a histogram of the raw data, a Gaussian mixture model (GMM) fitted to the data with the number of components (individual Gaussian distributions corresponding to each copy number) of the model determined by inspecting the number of peaks in the histogram, and by previous knowledge of the range of copy number variation in cohorts described previously. The variance of each Gaussian distribution was assumed to be the same when fitting the models to the data. Fitting the GMM allows integer copy numbers to be called from the data with an associated Bayesian posterior probability value for each call. It also allows fitting of two different models for cases and controls to provide a formal test of association of copy number with disease<sup>25</sup>. This analysis is implemented in the R package CNVtools v 1.42.3. Examples of GMM fits to the data are shown in supplementary figure 1, for PRT data.

The raw aCGH data for the Crohn's disease cohort was normalised two different ways; named as normalised1 and normalised2. In case of first normalisation (normalised1), the log of the ratio of the red and green channel data ( $\log(R/G)$ ) was used whereas in second normalisation (normalised2), the log of the ratio of the quantile normalised red and green channel data ( $\log(QNorm(R)/QNorm(G))$ ) was calculated. Data from 12 probes spanning CNV1 and 18 probes spanning CNV2 were summarised using the first principal component of the data, so that each sample had one summary value for CNV1 and one summary value for CNV2. For CNV1, plotting the data on a histogram gave three clusters and CNV1 copy number was called using a three-component Gaussian mixture model. For some samples, duplicate aCGH data was available. In such cases, the duplicate sample with the lowest posterior probability in support of an integer copy number was removed prior to case-control analysis. Normalised1 data were used for CNV1 case-control analysis, and normalised2 data were used for CNV2 analysis. Examples of Gaussian mixture model fits to the data are shown in supplementary figure 2, for aCGH data.

Fisher's exact tests and regression analyses were performed using R 3.1.0, and meta-analysis used the R package meta v4.3. All cohorts had a power of greater than 0.9 to detect the effect size previously observed<sup>13</sup> at a significance level of 0.05 or below, with the exception of the Danish cohort which had a power of 0.76.

## Results

We genotyped 1449 cases and 994 controls from the English, Scottish and Danish cohorts using our paralogue ratio test approach, described previously. The copy number distribution of CNV1 ranged from 0 to 4 in all populations, and CNV2 ranged from 1 to 11 (supplementary figures 1-2), which is consistent with previous studies on European populations<sup>20</sup>. Histograms of the raw data show clear clustering for CNV1 but poorer clustering for CNV2, where clear histogram peaks are seen only for lower copy numbers. Furthermore, visual inspection shows that for CNV2 calling quality varied from cohort to cohort, and this is reflected in the quality score of the Gaussian mixture model (Q, supplementary table 2) that is fitted to the data and used to call integer copy numbers. Because of this, we used two approaches for testing for association with disease. Firstly, we called integer copy number using CNVtools and used those copy numbers in a standard Fisher's exact test. Secondly, we used a feature of CNVtools that tests for association at the same time as fitting the Gaussian mixture model, which has the advantage of explicitly taking into account uncertainty in copy number calling.

A subset of both the English and Scottish cohorts had been analysed by array CGH as part of the Wellcome Trust Case Control Consortium genomewide CNV analysis. This allowed us to compare our copy number calling using PRT with DNA dosage data generated by array CGH. The arrayCGH data and PRT raw data were correlated for both CNV1 ( $r^2=0.75$  using normalised1 data,  $r^2=0.65$  using normalised2 data) and for CNV2 ( $r^2=0.55$  using normalised1 data,  $r^2=0.43$  using normalised2 data). Figure 2 shows that summarising the aCGH data as the first principal component of 12 probes spanning CNV1 gives concordance with PRT results, and suggests that these aCGH data could call CNV1 copy number quite robustly. Compared to the normalised2 approach (Figure 2b, see Methods), data normalised using the normalised1 approach showed a stronger correlation with the PRT raw data and showed better distinction between the two main peaks (copy numbers 1 and 2 (Figure 2a)). Therefore the normalised1 data were chosen for the full cohort analysis. For CNV2, although aCGH measures the DNA dosage and is correlated with PRT calls, there is a single continuous distribution with no evidence of clustering about integer copy numbers (Figure 3), and the correlation with PRT raw data is much weaker than is the case with the CNV1 data.

For the first test of association of copy number with CD, we followed the approach described in Renner et al. In that paper CNV1 was genotyped as the deletion *DMBT1*<sup>SR47-</sup> using a PCR-based approach. Previously, we have shown that this is a simplification of the CNV, with duplications also being observed in the population<sup>20</sup>. Copy number 0 is equivalent to a homozygous deletion *DMBT1*<sup>SR47-/-</sup>, 1 to a heterozygous deletion *DMBT1*<sup>SR47+/-</sup> and 2 to a homozygous reference *DMBT1*<sup>SR47+/+</sup>. Copy numbers 3 and 4 represent heterozygous and homozygous duplications

respectively. To directly compare our data with the previously published data, we called deletion genotype from our CNV1 data, grouping all CNV1 copy numbers of 2 and above as homozygous reference genotype. We also called deletion genotype from the WTCCC aCGH data from samples not included in our English and Scottish cohorts. For the 785 samples where we had matching PRT and array CGH data, 6 samples disagreed for the genotype called, giving a discordance rate of less than 1.6% (upper 95% confidence limit). All 785 samples with matching PRT data were removed from the WTCCC cohort analysis. Analysis of allele frequency counts in each cohort showed a higher frequency of the deletion allele in CD patients in three of the four cohorts, but the differences were not statistically significant (Mantel-Haenzel OR 1.10, 95%CI 0.97-1.24,  $p=0.40$ , Table 1, figure 4B).

We then asked whether full copy number typing of CNV1, where higher copy numbers corresponding to duplications can be called, would strengthen our association. Unfortunately, aCGH did not call high copy numbers effectively (figure 2, supplementary figure 2) and so we were limited to the cohorts typed by PRT. Using logistic regression to test the linear effect on CD case-control status with each increase in copy number, we found no significant effect (combined  $p=0.17$ , table 2, figure 3C), a result confirmed when analysed using the likelihood approach of CNVtools (supplementary table 2). This suggests that the duplication allele at CNV1 is unlikely to protect against CD, although given the low frequency of this allele we may not have power to detect anything but a strong effect.

We then examined whether copy number at CNV2 was associated with CD. Analysis of our three cohorts provided apparently contradictory results, with the Scottish cohort showing no association, the English cohort showing a marginally higher copy number ( $p=0.01$ ) in the CD patients and the Danish cohort showing a marginally lower ( $p=0.03$ ) copy number in the CD patients (table 3, supplementary table 3). This variation is due to variation in the patients rather than the controls, as the mean copy number in the controls is remarkably consistent across all three cohorts (table 3). The simplest interpretation of the results is that of stochastic variation about a null result, and indeed combining the datasets suggests this (combined  $p=0.446$ , Mantel-Haenzel OR =0.98, 95% CI 0.927-1.034)(Figure 3D). It may be the case that batch effects in typing high copy numbers of this CNV have generated this inconsistency. Indeed, even carefully designed CNV studies are prone to batch effects and the Scottish cohort was the only cohort where cases and controls originated from the same laboratory and were randomly distributed across all experimental plates.

For the Scottish cohort, age of CD first diagnosis data were available as a proxy for age of onset, and it is conceivable that CNV of the SRCR domains within *DMBT1* could affect this trait, notwithstanding an overall effect on risk of developing CD. We analysed the effect of copy number at both CNV1 and



CNV2 with age at diagnosis (table 4), controlling for the known effect of sex on age of onset. We confirmed that females have on average a later age of onset in this cohort, but found no evidence of an effect of CNV1 or CNV2. Analysis of CNV1 coded as *DMBT1*<sup>SR47</sup> genotype also showed no significant effect on age at diagnosis.

## Discussion

Previous work has shown the importance of *DMBT1* in the etiology of CD using studies of knockout mice and genetic association of the *DMBT1*<sup>SR47</sup>- deletion allele within the gene and CD. However, the genetic association had not been tested on another case-control cohort and had a relatively small sample size, and such association studies are prone to false positive results through differential bias or chance effects. Furthermore, the effect size observed (OR=1.75) is larger than most effect sizes identified by GWAS<sup>26</sup> and, if correct, could potentially be of clinical importance.

We conducted this study to try and replicate a previous genetic association study of the *DMBT1*<sup>SR47</sup>- deletion with CD. We used a combination of publically available data, generated as part of the Wellcome Trust Case Control Consortium study of copy number variation, and data we generated on three case-control cohorts using paralogue ratio tests to type the *DMBT1*<sup>SR47</sup>- deletion on a total of 2679 cases and 4088 controls. Comparisons between PRT raw data and array CGH data showed that while arrayCGH reflects copy number variation, correct normalisation is important to optimise the copy number calling, even when clear clusters of raw values are observed. After meta-analysis of our data, we did not replicate the original association<sup>13</sup>, and this could be due to a number of reasons. It is possible that, because we focused on Northern European populations and the original study was conducted on an Italian sample, the *DMBT1*<sup>SR47</sup>- deletion allele confers susceptibility to CD only in Italian populations, perhaps due to an interaction with diet. It is also possible that different diagnosis criteria were used, perhaps enriching for a particularly severe clinical phenotype in the original study, although there is no indication of this in the original study. However, the most likely explanation is that this was a false-positive result. It should be noted that in the original study the genotype frequencies for the cases show an extreme deviation from Hardy-Weinberg Equilibrium, with an excess of heterozygotes ( $p=5 \times 10^{-4}$ ,  $\chi^2$  test with 1 d.f.), which we do not observe in our data. We conducted a test for heterogeneity across our datasets and including the original data previously published, which suggested that the original dataset was from a distinct population ( $p=0.039$ ) and combining all the data in a meta-analysis would be inappropriate.

Previous analysis of CNVs within the *DMBT1* gene has shown that the *DMBT1*<sup>SR47-</sup> deletion is in fact part of a multiallelic CNV called CNV1, and that another CNV, called CNV2, is 3' to CNV1 and also affects the number of SRCR domains<sup>20</sup>. Using our PRT assays, we typed multiallelic copy number for both CNV1 and CNV2 on the Scottish, English and Danish case-control cohorts, and found no evidence of association. In this study, we use our copy number typing approaches to call the full spectrum of copy number variation at both CNV1 and CNV2. Given that the full range of copy numbers can be typed, we might expect more power to detect any association that was linearly dependent on copy number, but we do not detect such an effect for CNV1 nor CNV2, nor could we show any association with CNV1 copy number or CNV2 copy number.

One important feature of the *DMBT1*<sup>SR47-</sup> deletion allele is that, as part of CNV1, it has a remarkably high mutation rate of between 0.7%-2.7% per generation<sup>20</sup>. This has the consequence that *DMBT1*<sup>SR47-</sup> deletions will be generated by recurrent mutation, thereby eroding linkage disequilibrium with neighbouring SNP alleles. A recent study has identified a SNP allele associated with Crohn's disease within *DMBT1* at genomewide significance levels<sup>27</sup>. It is unclear why this allele has not been identified by GWAS studies, and indeed it may not be in LD with SNPs assayed by GWAS studies, so further research is needed to dissect the nature of this association. Our results in this study do not exclude an association of single nucleotide variation within *DMBT1* and CD, nor do they exclude a role for *DMBT1* in CD which has previously been suggested by the *Dmbt1* knockout mouse. Indeed, *DMBT1* shows increased expression in the intestinal mucosa in CD patients, and this increased expression is dependent on NOD2 activation, because this response is abolished in CD patients homozygous for a *NOD2* SNP allele causing a *NOD2* frameshift, an allele also associated with CD<sup>28</sup>. Given the role of *DMBT1* in binding bacteria<sup>29,30</sup>, it seems reasonable to assume that the *DMBT1*<sup>SR47-</sup> deletion allele encodes a protein that has an altered interaction with the intestinal flora, and mediates its effect via its interactions with bacteria. However, our study has excluded a role for the extensive copy number variation of *DMBT1* in strongly modifying the susceptibility to CD.

## Acknowledgements

This work was funded by a Government of India Ministry of Social Justice and Empowerment PhD studentship to SP and EJH. EJH was supported in part by an MRC New Investigator Grant (GO801123). This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from [www.wtccc.org.uk](http://www.wtccc.org.uk). Funding for the project was provided by the Wellcome Trust under award

297 076113, 085475 and 090355. We would like to thank Mark Jobling and Jenny Bowdrey for access and  
298 support in using an ABI 3130xl capillary electrophoresis machine.

299 **Conflict of Interest**

300 None to declare.

## 301 Figure legends

### 302 Figure 1 Overview of the copy number variation at *DMBT1*

303 A dotplot shows the repeated nature of the *DMBT1* gene (shown from a screenshot from the UCSC  
304 genome browser). The tandemly-arranged SRCR repeat regions are shown, including SRCR14 which  
305 does not bind bacteria. The genome assembly shows one assembled copy of CNV1 and four  
306 assembled copies of CNV2. CNV regions, as recorded in the Database of Genome Variants, are  
307 shown below the *DMBT1* gene structure. Below these, location of reference and test amplicons of  
308 the four independent paralogue ratio tests (PRTs) that measure copy number of CNV1 and CNV2 are  
309 shown.

### 310 Figure 2 Analysis of calling CNV1 copy number using PRT and arrayCGH

311 a) 688 samples from the English CD cohort and 97 samples from the Scottish CD cohort with copy  
312 number measured by both PRT (y axis) and aCGH (x axis). aCGH data here is normalised using  
313  $(\log(R/G))$  and represents the first principal component value of 12 probes.

314 b) as above but with aCGH data here normalised using  $(\log(QNorm(R)/QNorm(G)))$  where Qnorm is  
315 quantile-normalised.

### 316 Figure 3 Analysis of calling CNV2 copy number using PRT and arrayCGH

317 a) 688 samples from the English CD cohort and 97 samples from the Scottish CD cohort with copy  
318 number measured by both PRT (y axis) and aCGH (x axis). aCGH data here is normalised using  
319  $(\log(R/G))$  and represents the first principal component value of 18 probes.

320 b) as above but with aCGH data here normalised using  $(\log(QNorm(R)/QNorm(G)))$  where Qnorm is  
321 quantile-normalised.

322

### 323 Figure 4 Meta-analysis of cohorts in the association study

324 Forest plots of odds ratios for the deletion variant of *DMBT1*<sup>SR47</sup> for Scottish, English, Danish and  
325 WTCCC datasets only (a) Forest plots of the odds ratio per copy for CNV1 (b) and CNV2 (c). Each  
326 diagram displays the odds ratios for each dataset as a box with the 95% confidence interval marked  
327 by lines. The “MH Summary” represents the 95% confidence interval of the Mantel-Haenszel  
328 combined odds ratio for all datasets, whereas “Combined” represents the 95% confidence interval  
329 for totals for CNV1 and CNV2.

330

331 **Supplementary figure 1 Quality of CNV clustering by PRT**

332 This shows, for each cohort and CNV, a histogram of PRT copy number values (x-axis), normalised so  
333 that the entire distribution has a standard deviation of 1, to optimise Gaussian mixture model fitting.  
334 The lines show the fitted Gaussian mixture model separating each peak in the histogram.

335 **Supplementary figure 2 Quality of CNV1 clustering on the WTCCC cohort**

336 This shows, for three WTCCC cohorts, a histogram of the first principal component of 12 aCGH  
337 probes, normalised using the normalised1 method. The lines show the fitted Gaussian mixture  
338 model separating each peak in the histogram. Note that the direction of signal of the first principal  
339 component is arbitrary, so for the Crohn's disease cases the histogram is reversed, as compared to  
340 the control sample histograms.

341

342

343 **Tables**344 **Table 1 Association analysis of *DMBT1*<sup>SR47</sup> genotype with Crohn's disease**

345

| Population   | Cohort          | <i>DMBT1</i> <sup>SR47-/-</sup><br>Number<br>(frequency) | <i>DMBT1</i> <sup>SR47+/-</sup><br>Number<br>(frequency) | <i>DMBT1</i> <sup>SR47+/+</sup><br>Number<br>(frequency) | total       | Fisher's<br>Exact<br>Test p<br>value | Odds Ratio                   |
|--------------|-----------------|--|--|--|-------------|--------------------------------------|------------------------------|
| Scottish     | CD              | 5 (0.01)   | 57 (0.16)  | 286 (0.82)   | 348         | 0.93                                 | 0.97 (0.68-<br>1.39)         |
|              | controls        | 3 (0.01)   | 61 (0.18)  | 276 (0.81)   | 340         |                                      |                              |
| English      | CD              | 7 (0.01)   | 178 (0.19)   | 761 (0.80)   | 946         | 0.20                                 | 1.19 (0.91-<br>1.56)         |
|              | controls        | 2 (0.00)   | 79 (0.16)  | 399 (0.83)   | 480         |                                      |                              |
| Danish       | CD              | 5 (0.03)   | 34 (0.22)  | 116 (0.75)   | 155         | 0.09                                 | 1.53 (0.95-<br>2.46)         |
|              | controls        | 4 (0.02)   | 26 (0.15)  | 144 (0.83)   | 174         |                                      |                              |
| WTCCC        | CD              | 16 (0.01)  | 226 (0.18)   | 988 (0.80)   | 1230        | 0.76                                 | 1.05 (0.90-<br>1.23)         |
|              | controls        | 41 (0.01)  | 535 (0.17)   | 2517 (0.81)  | 3093        |                                      |                              |
| <b>Total</b> | <b>CD</b>       | <b>33 (0.01)</b>   | <b>495 (0.19)</b>  | <b>2151 (0.80)</b>                                       | <b>2679</b> | <b>0.22</b>                          | <b>1.07 (0.96-<br/>1.21)</b> |
|              | <b>controls</b> | <b>50 (0.01)</b>   | <b>701 (0.17)</b>  | <b>3337 (0.82)</b>                                       | <b>4087</b> |                                      |                              |

346

347

348 **Table 2 Association analysis of *DMBT1* CNV1 copy number with Crohn's disease**

|                      | Scottish            |          | English             |          | Danish              |          |
|----------------------|---------------------|----------|---------------------|----------|---------------------|----------|
| CNV1 Copy number     | CD                  | controls | CD                  | controls | CD                  | controls |
| 0                    | 5                   | 3        | 7                   | 2        | 5                   | 4        |
| 1                    | 57                  | 61       | 178                 | 79       | 34                  | 26       |
| 2                    | 275                 | 263      | 731                 | 387      | 114                 | 139      |
| 3                    | 11                  | 12       | 30                  | 11       | 2                   | 5        |
| 4                    | 0                   | 1        | 0                   | 1        | 0                   | 0        |
| n                    | 348                 | 340      | 946                 | 480      | 155                 | 174      |
| failed               | 2                   | 0        | 3                   | 0        | 1                   | 5        |
| mean                 | 1.84                | 1.84     | 1.83                | 1.85     | 1.73                | 1.83     |
| sd                   | 0.48                | 0.49     | 0.47                | 0.44     | 0.54                | 0.49     |
| OR (95% CI) per copy | 0.979 (0.717-1.335) |          | 0.886 (0.694-1.126) |          | 0.673 (0.435-1.028) |          |
| P (log reg)          | 0.891               |          | 0.323               |          | 0.070               |          |

349

350

351 **Table 3 Association analysis of *DMBT1* CNV2 copy number with Crohn's disease**

|                        | Scottish         |          | English             |          | Danish              |          |
|------------------------|------------------|----------|---------------------|----------|---------------------|----------|
| CNV2<br>Copy<br>number | CD               | controls | CD                  | controls | CD                  | controls |
| 1                      | 0                | 0        | 1                   | 0        | 0                   | 0        |
| 2                      | 13               | 8        | 14                  | 11       | 2                   | 9        |
| 3                      | 29               | 34       | 98                  | 48       | 12                  | 21       |
| 4                      | 82               | 79       | 229                 | 110      | 30                  | 44       |
| 5                      | 78               | 83       | 319                 | 125      | 46                  | 43       |
| 6                      | 70               | 69       | 164                 | 104      | 36                  | 33       |
| 7                      | 37               | 26       | 67                  | 51       | 16                  | 11       |
| 8                      | 23               | 18       | 35                  | 23       | 10                  | 11       |
| 9                      | 11               | 10       | 3                   | 4        | 4                   | 5        |
| 10                     | 2                | 1        | 2                   | 4        | 0                   | 0        |
| 11                     | 1                | 0        | 0                   | 0        | 0                   | 0        |
| n                      | 346              | 328      | 932                 | 480      | 156                 | 177      |
| failed                 | 4                | 12       | 17                  | 0        | 0                   | 2        |
| mean                   | 5.27             | 5.15     | 4.95                | 5.15     | 5.34                | 4.97     |
| sd                     | 1.68             | 1.56     | 1.31                | 1.49     | 1.46                | 1.63     |
| OR                     | 1.05 (0.95-1.15) |          | 0.901 (0.832-0.976) |          | 1.169 (1.017-1.348) |          |
| p                      | 0.329            |          | 0.0103              |          | 0.0297              |          |

352

353

354



355 **Table 4** Multiple linear regression analysis testing association of CNV with age of onset

| Variable                  | B      | B (Standard error) | t-statistic | P                    |
|---------------------------|--------|--------------------|-------------|----------------------|
| (Intercept)               | 30.4   | 7.31               | 4.15        | $4.3 \times 10^{-5}$ |
| Sex<br>(reference=female) | -4.76  | 1.83               | -2.61       | 0.0097               |
| CNV1                      | 0.804  | 1.62               | 0.43        | 0.67                 |
| CNV2                      | 0.0113 | 0.513              | 0.022       | 0.98                 |

356 N=306, 7 omitted due to missing data

357 Predictor variables were sex, CNV1 copy number and CNV2 copy number, with age at diagnosis the  
 358 dependent variable. The values for the effect size (B) with its standard error are given, together with  
 359 the corresponding t-statistic used to test whether the value of B is significantly different from zero.  
 360 The p value of that test is given in the rightmost column.

361

362

363

## References

1. Stone MA, Mayberry JF, Baker R: Prevalence and management of inflammatory bowel disease: a cross-sectional study from central England. *European journal of gastroenterology & hepatology* 2003; **15**: 1275-1280.
2. So HC, Gui AH, Cherny SS, Sham PC: Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. *Genetic epidemiology* 2011; **35**: 310-317.
3. Franke A, McGovern DP, Barrett JC *et al*: Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics* 2010; **42**: 1118-1125.
4. Jostins L, Ripke S, Weersma RK *et al*: Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 2012; **491**: 119-124.
5. Wain LV, Armour JAL, Tobin MD: Genomic copy number variation, human health, and disease. *The Lancet* 2009; **374**: 340-350.
6. Falchi M, Moustafa JSE-S, Takousis P *et al*: Low copy number of the salivary amylase gene predisposes to obesity. *Nature Genetics* 2014; **46**: 492-497.
7. Hollox EJ, Huffmeier U, Zeeuwen PL *et al*: Psoriasis is associated with increased beta-defensin genomic copy number. *Nature Genetics* 2008; **40**: 23.
8. Craddock N, Hurles ME, Cardin N *et al*: Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* 2010; **464**: 713-720.
9. Fellermann K, Stange DE, Schaeffeler E *et al*: A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *The American Journal of Human Genetics* 2006; **79**: 439-448.
10. Bentley RW, Pearson J, Gearry RB *et al*: Association of higher DEFB4 genomic copy number with Crohn's disease. *The American journal of gastroenterology* 2009; **105**: 354-359.
11. Aldhous MC, Bakar SA, Prescott NJ *et al*: Measurement methods and accuracy in copy number variation: failure to replicate associations of beta-defensin copy number with Crohn's disease. *Human molecular genetics* 2010; **19**: 4930-4938.

- 404 12. Fode P, Jespersgaard C, Hardwick RJ *et al*: Determination of beta-defensin genomic copy  
405 number in different populations: a comparison of three methods. *PLOS one* 2011; **6**: e16768.
- 406
- 407 13. Renner M, Bergmann G, Krebs I *et al*: < i> DMBT1</i> Confers Mucosal Protection In Vivo  
408 and a Deletion Variant Is Associated With Crohn's Disease. *Gastroenterology* 2007; **133**:  
409 1499-1509.
- 410
- 411 14. Madsen J, Mollenhauer J, Holmskov U: Review: Gp-340/DMBT1 in mucosal innate immunity.  
412 *Innate immunity* 2010; **16**: 160-167.
- 413
- 414 15. Bikker FJ, Ligtenberg AJ, End C *et al*: Bacteria binding by DMBT1/SAG/gp-340 is confined to  
415 the VEVLXXXXW motif in its scavenger receptor cysteine-rich domains. *Journal of Biological*  
416 *Chemistry* 2004; **279**: 47699-47703.
- 417
- 418 16. Bikker FJ, Ligtenberg AJ, Nazmi K *et al*: Identification of the bacteria-binding peptide domain  
419 on salivary agglutinin (gp-340/DMBT1), a member of the scavenger receptor cysteine-rich  
420 superfamily. *Journal of Biological Chemistry* 2002; **277**: 32109-32115.
- 421
- 422 17. De Lisle RC, Xu W, Roe BA, Ziemer D: Effects of Muclin (Dmbt1) deficiency on the  
423 gastrointestinal system. *American Journal of Physiology-Gastrointestinal and Liver*  
424 *Physiology* 2008; **294**: G717-G727.
- 425
- 426 18. Mollenhauer J, Wiemann S, Scheurlen W *et al*: DMBT1, a new member of the SRCR  
427 superfamily, on chromosome 10q25. 3–26.1 is deleted in malignant brain tumours. *Nature*  
428 *Genetics* 1997; **17**: 32-39.
- 429
- 430 19. Sasaki H, Betensky RA, Cairncross JG, Louis DN: DMBT1 Polymorphisms Relationship to  
431 Malignant Glioma Tumorigenesis. *Cancer research* 2002; **62**: 1790-1796.
- 432
- 433 20. Polley S, Louzada S, Forni D *et al*: Evolution of the rapidly-mutating human salivary agglutinin  
434 gene (DMBT1) and population subsistence strategy. *Proceedings of the National Academy of*  
435 *Sciences* 2015; **112**: 5105-5110.
- 436
- 437 21. Jespersgaard C, Fode P, Dybdahl M *et al*: Alpha-defensin DEFA1A3 gene copy number  
438 elevation in Danish Crohn's disease patients. *Digestive diseases and sciences* 2011; **56**: 3517-  
439 3524.
- 440
- 441 22. Vind I, Riis L, Jespersgaard C *et al*: Genetic and environmental factors as predictors of disease  
442 severity and extent at time of diagnosis in an inception cohort of inflammatory bowel  
443 disease, Copenhagen County and City 2003–2005. *Journal of Crohn's and Colitis* 2008; **2**:  
444 162-169.
- 445

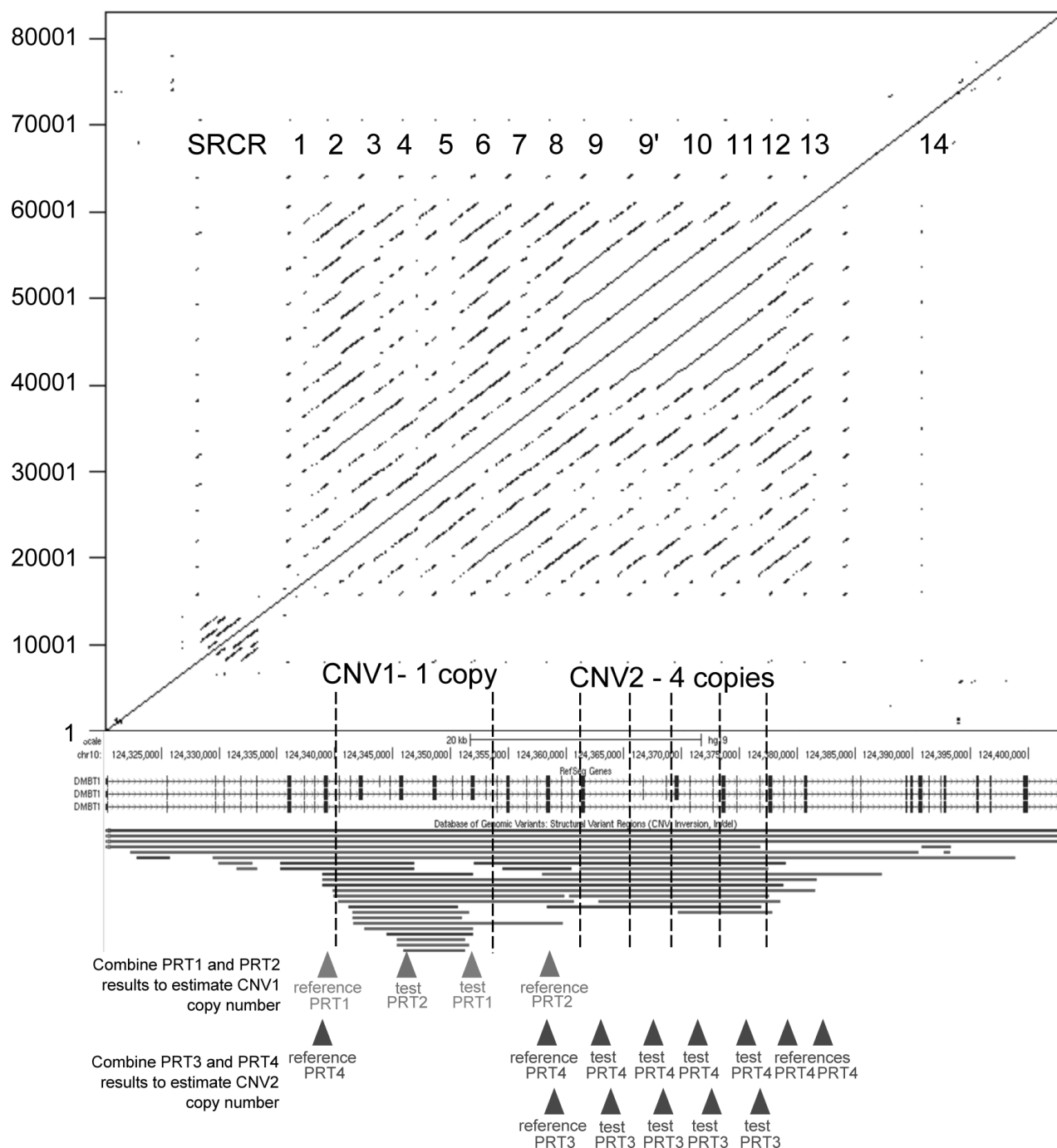
- 446 23. Prescott NJ, Fisher SA, Franke A *et al*: A Nonsynonymous SNP in *ATG16L1* Predisposes  
 447 to Ileal Crohn's Disease and Is Independent of *CARD15* and *IBD5*.  
 448 *Gastroenterology* 2007; **132**: 1665-1671.
- 449  
 450 24. Armour JAL, Palla R, Zeeuwen PLJM, den Heijer M, Schalkwijk J, Hollox EJ: Accurate, high-  
 451 throughput typing of copy number variation using paralogue ratios from dispersed repeats.  
 452 *Nucleic acids research* 2007; **35**: e19-e19.
- 453  
 454 25. Barnes C, Plagnol V, Fitzgerald T *et al*: A robust statistical method for case-control  
 455 association testing with copy number variation. *Nature Genetics* 2008; **40**: 1245-1252.
- 456  
 457 26. Liu JZ, Anderson CA: Genetic studies of Crohn's disease: past, present and future. *Best*  
 458 *Practice & Research Clinical Gastroenterology* 2014.
- 459  
 460 27. Diegelmann J, Czamara D, Le Bras E *et al*: Intestinal DMBT1 Expression Is Modulated by  
 461 Crohn's Disease-Associated IL23R Variants and by a DMBT1 Variant Which Influences Binding  
 462 of the Transcription Factors CREB1 and ATF-2. *PLOS one* 2013; **8**: e77773.
- 463  
 464 28. Rosenstiel P, Sina C, End C *et al*: Regulation of DMBT1 via NOD2 and TLR4 in intestinal  
 465 epithelial cells modulates bacterial recognition and invasion. *The Journal of Immunology*  
 466 2007; **178**: 8203-8211.
- 467  
 468 29. Loimaranta V, Hytönen J, Pulliainen AT *et al*: Leucine-rich repeats of bacterial surface  
 469 proteins serve as common pattern recognition motifs of human scavenger receptor gp340.  
 470 *Journal of Biological Chemistry* 2009; **284**: 18614-18623.
- 471  
 472 30. Madsen J, Tornøe I, Nielsen O *et al*: CRP-ductin, the mouse homologue of gp-340/deleted in  
 473 malignant brain tumors 1 (DMBT1), binds gram-positive and gram-negative bacteria and  
 474 interacts with lung surfactant protein D. *European journal of immunology* 2003; **33**: 2327-  
 475 2336.

476

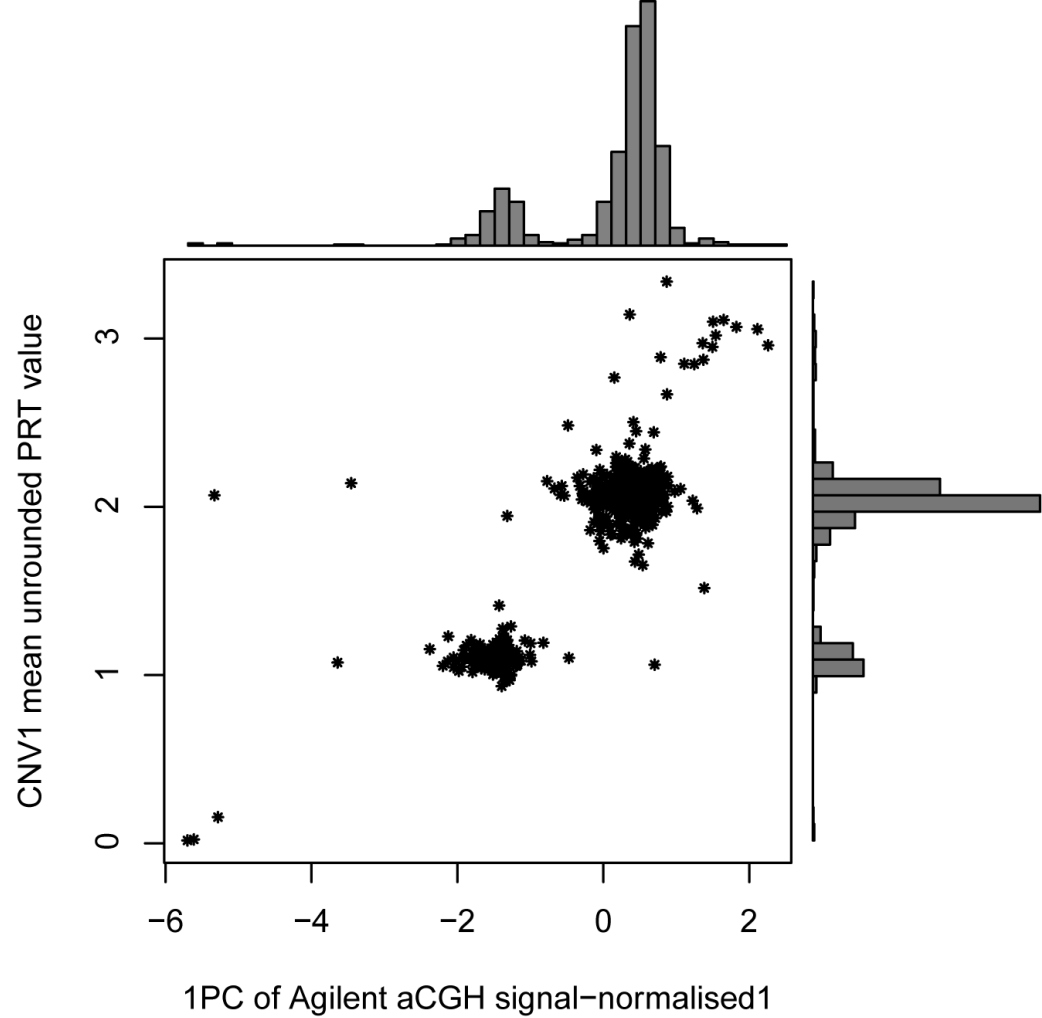
477

478

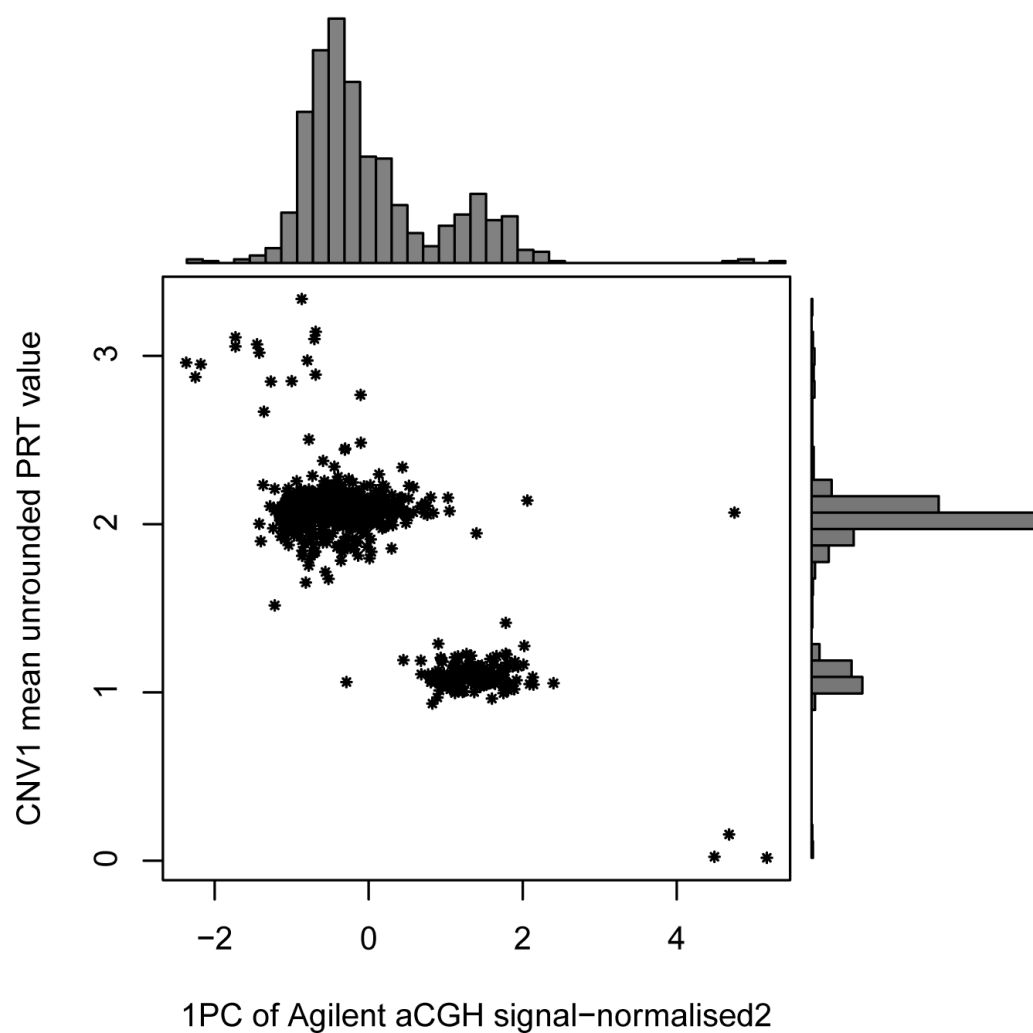
chr10:124320181-124403252



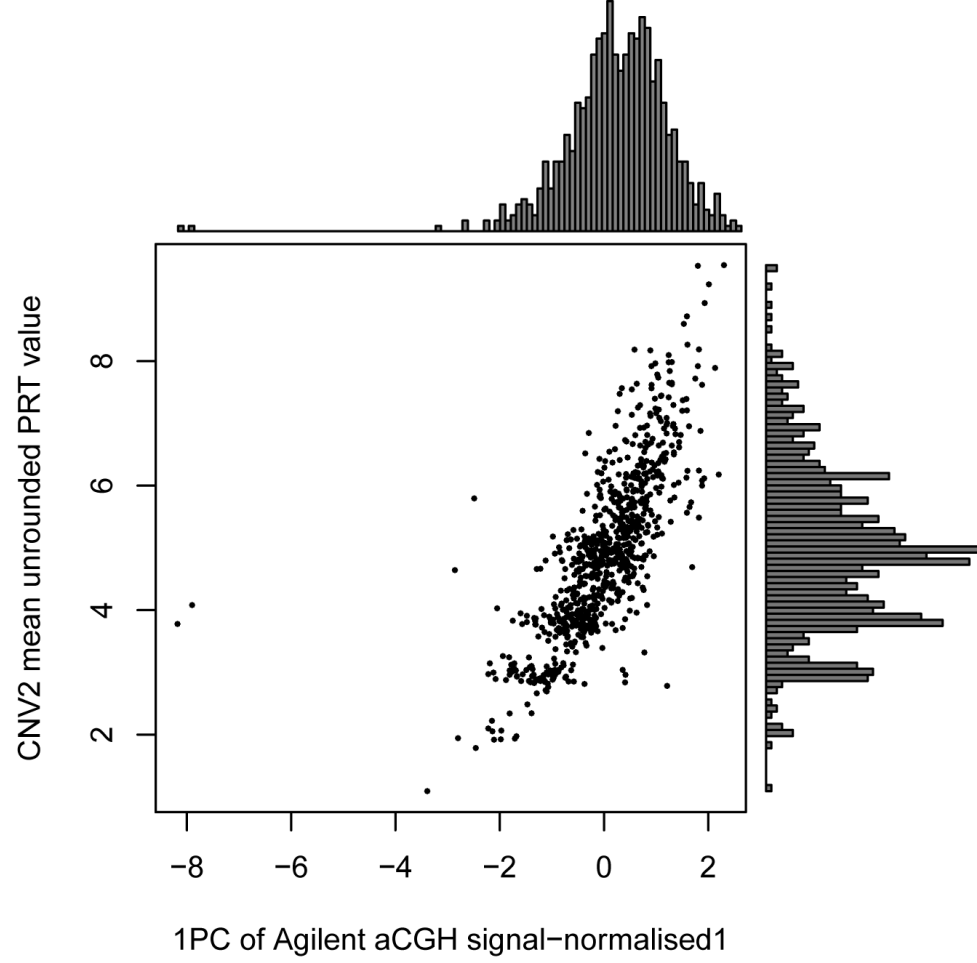
a)



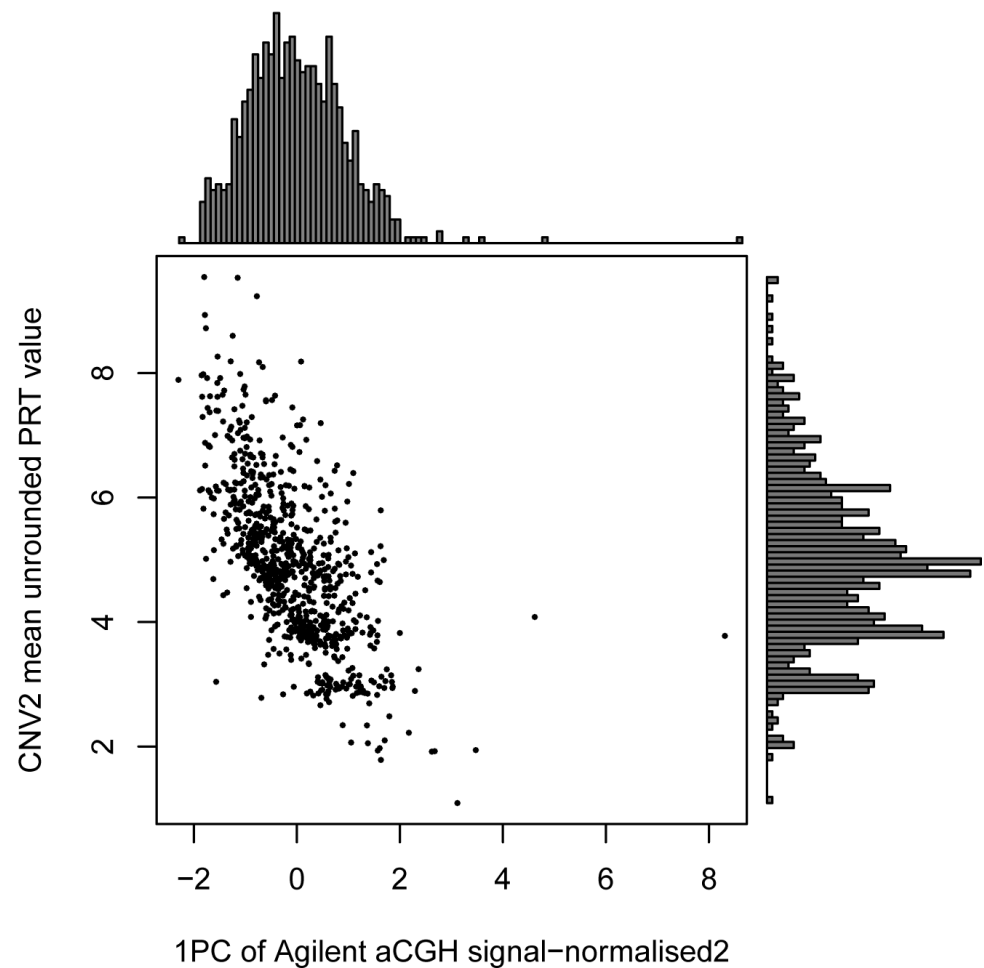
b)



a)



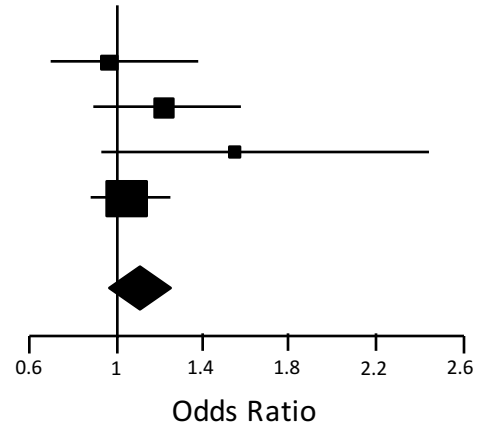
b)



A.

*DMBT1*<sup>SR47</sup>

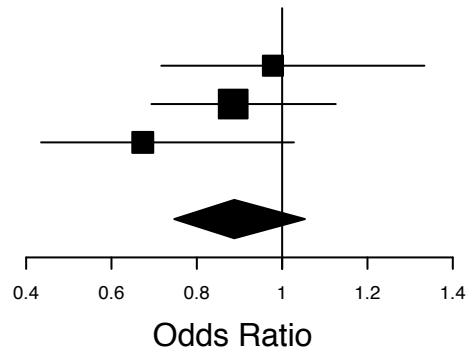
| Study             | OR           |
|-------------------|--------------|
| Scottish          | 0.975        |
| English           | 1.193        |
| Danish            | 1.527        |
| WTCCC             | 1.053        |
| <b>MH Summary</b> | <b>1.097</b> |



B.

CNV1

| Study           | OR           |
|-----------------|--------------|
| Scottish        | 0.988        |
| English         | 0.885        |
| Danish          | 0.673        |
| <b>Combined</b> | <b>0.888</b> |



C.

CNV2

| Study           | OR           |
|-----------------|--------------|
| Scottish        | 1.048        |
| English         | 0.901        |
| Danish          | 1.169        |
| <b>Combined</b> | <b>0.979</b> |

