Experimental aspects of copy number variant assays at *CCL3L1*

Sarah F. Field[1*], Joanna M. M. Howson[1*], Lisa M. Maier[2,3*], Susan Walker[4], Neil M Walker[1], Deborah J. Smyth[1], John A. L. Armour[4], David G. Clayton[1], John A. Todd[1]

[*] Authors contributed equally

[1] Juvenile Diabetes Research Foundation / Wellcome Trust Diabetes and Inflammation Laboratory, Cambridge Institute for Medical Research, Department of Medical Genetics, University of Cambridge, UK

[2]Division of Molecular Immunology, Center for Neurologic Diseases, Brigham and Women's Hospital and Harvard Medical School, Boston, USA

[3]Program in Medical and Population Genetics, Broad Institute, Massachusetts Institute of Technology and Harvard University, Cambridge, USA

[4]Institute of Genetics and School of Biology, University of Nottingham, Queen's Medical Centre, Nottingham, UK

Supplementary Methods

Variants of CCL3L1 and CCL4L1

There are three known SNPs in *CCL3L1* and eleven in *CCL4L1* that define variants of these genes, known as *CCL3L3* and *CCL4L2*, respectively. Neither the qPCR assay nor any of the PRT assays are capable of distinguishing these variants. Therefore, the copy numbers that are reported should be assumed to be for both variants of each gene (**Supplementary Figure 1**).

Subjects

In total, 5,771 British T1D cases and 6,854 British controls were studied. The T1D cases were collected as part of the Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory's Genetic Resource Investigating Diabetes (GRID) study (http://www-gene.cimr.cam.ac.uk/todd/). All T1D cases were under age 17 years at diagnosis. The controls were from the British 1958 Birth Cohort (http://www.b58cgene.sgul.ac.uk/index.php), and were matched to the cases using place of recruitment, for each of 12 geographical regions across Great Britain (Southern England, South-western England, South-eastern England, Eastern England, London, Midlands, Wales, North-eastern England, North Midlands, East and West Ridings, Northern England and Scotland) to minimise bias in our association results owing to varying disease prevalence and allele frequencies across Great Britain[1]. All cases and controls were of self-reported white ethnicity. 4,646 T1D cases and 4,989 controls were studied using qPCR and 4,910 T1D cases and 5,046 controls were studied using the PRT assay. 3,785 cases and 3,181 controls were common to both experiments. All DNA samples were collected with approval from the Cambridgeshire 2 Research Ethics Committee, and written consent was obtained from all participants, or parents of participants who were too young to consent.

PRT

The PRT [2] relies on finding a locus paralogous to the CNV locus with invariant copy number *e.g.* has 2 copies per diploid genome. A pair of primers is designed which binds both at the CNV locus and at a paralogous reference locus, but which produces fragments of different lengths for the CNV and reference locus. For the *CCL3L1* region Walker *et al.* developed three assays[3], using all available consensus sequences to design the primers and estimate PCR amplicon size (NCBI (http://www.ncbi.nlm.nih.gov), UCSC (http://genome.ucsc.edu), VEGA (http://vega.sanger.ac.uk), and HapMap (http://www.hapmap.org)). The first measures *CCL3L1* copy number by using *CCL3* as the reference locus and *CCL3L1* as the CNV locus; the primer pair (TCATAGTGGGTTCTCTGTTTC-Forward and ATCCAGGGCTGCTTACTT-Reverse) amplifies a 220 bp fragment in *CCL3L1* and a 226 bp fragment in *CCL3*. The second measures *CCL4L1* copy number by using *CCL4* as the reference locus and *CCL4L1* is the CNV; the primer pair (GAGTCTGCTTCCAGTGCT-Forward and GAGGAGTCCTGAGTATGGAG-Reverse) amplifies a 340 bp fragment in *CCL4L1* and a 326 bp fragment in *CCL4*. The third measures a LTR sequence that lies between *CCL3L1* and *CCL4L1* as the CNV locus and a paralogous, non-duplicated, LTR on chromosome 10q22 as the reference locus, such that this ratio provides an indirect measure of the *CCL3L1* and *CCL4L1* region copy number; the primer pair (AGTTTTCCTCTGCCTAGC-Forward and TATTTATTTTAAGGTGTGCAC-Reverse) amplifies a 368 bp LTR fragment in the *CCL3L1* CNV region on 17q12 and a 377 bp LTR fragment on chromosome 10q22. The products were amplified in a 10 µl reaction using 5 ng of template DNA, the primer concentration was 500 nM, the reaction was run for 25 PCR cycles. The ratio of the heights of the peaks corresponding to the two fragments (test and reference) for each assay was multiplied by 2 (as the reference fragment is assumed to have two copies) to give the unscored copy number, referred to as the assay ratio. Whilst Walker *et al.* [3]advocated averaging across all three assays to obtain the number of copies of *CCL3L1*, we were interested in the performance of each assay and so analysed the ratios from each assay individually in addition to averaging across assays (**Supplementary Results**). In order to minimise biased associations due to plate-to-plate variation in PCR quality, cases and controls were assayed on the same 96 well plates with the scientist blind to case/control status, such that each plate contained 45 cases and 45 controls. In addition, each 96 well plate included six standard samples with known copy number (four CEPH samples, one T1D case and the A431 cell line: CEPH1334.01 = 0 copies, CEPH1334.10 = 1 copy, CEPH1375.11 = 2 copies, A431 = 2 copies, CEPH1334.12 = 3 copies and 101732899M = 4 copies). These samples were included for quality control (QC) purposes. The PCR products for each dye were pooled within the LTR, *CCL4L1* and *CCL3L1* assays individually and 2 μl of the pooled product was analysed by capillary electrophoresis on a 3730 AB sequencer. The electropherograms of the PRT assays showed easily distinguishable peaks with a centre point that did not vary by more than one bp. We observed differences in read quality with different dyes, and hence the LTR assay and the *CCL4L1* assay were both performed in triplicate using FAM, HEX and ROX dyes. The *CCL3L1* assay was performed in duplicate using HEX and ROX (FAM was attempted, but consistently produced an artefact peak at 226 bp so an accurate measurement was not possible). The whole experiment was performed in duplicate such that each sample would appear on two plates with one plate arbitrarily named the "original" and the other plate arbitrarily named the "replicate".

Quality Control procedures: PRT

The LTR on chromosome 17q12 generally gives PCR products of 368 bp but with a 364 bp variant. When present, the 364 bp variant amplifies with approximately 23% greater yield per target copy, and peak heights for the 364 bp peak were therefore scaled by a factor of 0.77 before further analysis.

As measured by capillary electrophoresis, all samples for which *CCL4* had a peak and *CCL4L1* did not, were assumed to correspond to zero copies of *CCL4L1*. Similarly all samples for which *CCL3* had a peak and *CCL3L1* did not, were assumed to correspond to zero copies of *CCL3L1* and all samples for which LTR10 had a peak and LTR17 did not were also assumed to correspond to zero copies of the CNV. The test sample CEPH1334.01, was correctly called as a zero on all 104 plates on which it passed QC.

Initially we analysed original plates and replicate plates separately. So for each PRT assay across original plates and separately across replicate plates, the ratio of the height of the peak of the CNV (*CCL3L1*, *CCL4L1* or LTR17) to the height of the peak of the reference locus (*CCL3*, *CCL4* or LTR10) was calculated for each of the dyes. Inconsistencies in the ratios across the dyes were considered. Those samples with a maximum difference between ratios greater than 3 standard deviations (SD) from the mean maximum difference between ratios across all plates, were rejected (again this was done separately for original and replicate plates). On the original plates, 153 samples were dropped from the *CCL4L1* assay, 155 samples were dropped from the *CCL3L1* assay and 131 samples were dropped from the LTR assay, similarly on the replicate plates, 168 samples were dropped for the *CCL4L1* assay, 137 samples were dropped from the *CCL3L1* assay and 138 samples were dropped from the LTR assay.

In order to ascertain that PCR failure rates did not introduce differential bias, we examined the samples that failed on the "original" plates, but passed on the duplicate, and these were found to form a distribution that closely followed that of the whole distribution. This was also true of the samples that passed QC on the "original" plate but failed on the "replicate" plate (**Supplementary Table 5**).

Having calculated unscored copy number (assay ratios) as twice the average of the raw ratios across dyes, we compared the absolute difference in copy number generated on the original and replicate plates. The *CCL3L1* assay was compared for original and replicate plates and the ones deviating by more than 3 SD from the mean difference in copy number across all plates were rejected (230 samples). We applied the same procedure to the *CCL4L1* and LTR assays and rejected 218 and 178 samples, respectively. In total 4,316 T1D case samples and 4,568 control samples were successful on at least one of the LTR, *CCL3L1* or *CCL4L1* assays.

Six samples known to have copy numbers between 0 and 4 were included on all plates for QC purposes. Having calculated integer copy numbers from the assay ratios, we found the majority of QC replicate samples had consistent copy numbers for the k-means clustered LTR assay data. Using the LTR copy number (assigned by k-means) the zero copy sample, CEPH1334.01, was correctly called as a zero, the one-copy sample, CEPH1334.10, was correctly called as one, and the two-copy sample, CEPH1375.11, was correctly called as a two across all replicates. The greatest inconsistency was observed with the three-copy sample, CEPH1334.12 where 7% of replicates were incorrectly called as two copies (this was true of the rounded data also). The four-copy sample, 101732899M, had just one misclassification across the 73 replicates. Hence, the consistency was very high, with 98.6% of all replicated QC samples called correctly. This compares to 95.2% with the *CCL3L1* copy number (assigned by k-means) and 98.9% with the *CCL4L1* copy number (assigned by k-means). The k-means clustered data had much higher consistency across replicates than the rounded data, which was as low as 75% for *CCL3L1* and 81% for *CCL4L1*.

We note that the A431 cell line was consistently called as two copies with the *CCL3L1* and LTR assays using k-means clustering, whereas the *CCL4L1* assay was called as one copy, suggesting that *CCL4L1* and *CCL3L1* have inconsistent copy number in this cell line.

qPCR

Genotyping was performed according to the method of Gonzalez *et al.*[4]. Using published probes and primers (Gonzalez *et al.*[4]), quantitative real-time PCR (qPCR) (TaqMan) was performed using an AB7900 sequence detector system (Applied Biosystems). Cases and controls were aliquoted onto different plates. Reactions were performed in 384-well plate format, in a 5 µl reaction using 4 ng of template DNA (as determined by duplicate picogreen assay), the primer concentration in the reaction was 900 nM and the probe concentration was 250 nM. Duplicate wells were set up for *CCL3L1* and the reference locus hemoglobin beta (*HBB*) reactions, *i.e.* for each sample there were two measures of *CCL3L1* and two measures of *HBB* on a given plate. Each 384-well plate was in turn duplicated (so there were four measures of *CCL3L1* in total for each sample and four measures of *HBB*). We arbitrarily named one plate the “original” plate, and the second, the “replicate plate”. The threshold cycle (CT), which reflects the number of cycles at which the fluorescence generated within a reaction crosses a given threshold, was determined for each sample. Reactions for serial 1:2 dilutions (starting from 20 ng) of genomic DNA from A431 cells were performed on each 384-well plate and used as the standard curve. The standard curve was used to calibrate DNA concentration between *HBB* and the *CCL3L1* PCR reactions as they were in different wells on the plate.

Quality Control procedures: qPCR

All data analyses were performed within STATA (www.stata.com) using routines developed in-house (www-gene.cimr.cam.ac.uk/clayton/software).

A number of QC procedures were applied to the data generated using the qPCR assay. Firstly, all samples with fewer than four data points on the original plate or on the replicate plate were removed (n=138). We next calculated the difference in CT between samples duplicated on the *same* plate for both *CCL3L1* and *HBB*. This was done separately for the original plates and the replicate plates (as if the experiments were independent). Those samples with a mean difference in CT greater than three standard deviations (SD) away from the overall mean difference in CT across all original plates (or similarly across all replicate plates) were dropped. Note, this rejection criteria was applied to all samples on the original plates based on data obtained from all original plates and not on a plate by plate basis and likewise for the replicate plates, with the intention that all plates were treated consistently. Template quantity of each sample was calculated using standard curves on each plate, by regressing CT against log of the concentration of the A431 DNA for both *CCL3L1* and *HBB*. The *CCL3L1* copy number was calculated as twice the ratio of the template quantity of *CCL3L1* to the template quantity of *HBB*.

Across all "original plates" 120 samples had inconsistent CT for *CCL3L1* and 105 samples had inconsistent CT for *HBB* (*i.e.* the difference between the original and replicate sample was greater than 3 SD from the mean difference). Six of these had both *CCL3L1* and *HBB* inconsistencies on the "original" plates. Across all "replicate" plates, 105 samples had inconsistent CT for *CCL3L1* and 62 samples had inconsistent CT for *HBB*. Five of these had both *CCL3L1* and *HBB* inconsistencies on the "replicate" plates. One plate of control samples had elevated inconsistency between replicates (within plate) and so was dropped due to concerns regarding possible aliquoting errors or DNA quality. Three plates of T1D case samples had inconsistent standard curve copy number counts between the original plate and the replicated plate and so these plates were also dropped.

CCL3L1 copy numbers were calculated for both original and replicate plates as described above and, by Gonzalez *et al.*[4] as twice the ratio of *CCL3L1* to *HBB*. Consistency between original and replicate plates was checked by calculating the difference between the copy number on the original and replicate plates. Those differing by more than 3 SD from the mean copy number difference were removed. We noted that reproducibility for these values was lower in the higher copy numbers.

In contrast to the average of all CNV values passing the QC criteria, 2.41, the average of those values failing was 4.28, indicating that the performance of the assay for estimating higher copy numbers was not as good as for smaller copy numbers. In total 3,550 T1D case samples and 4,568 control samples were successfully genotyped on either the original or replicate plate, giving 2,608 cases and 2,513 controls common to both the PRT and qPCR experiments.

Statistics

Scoring of assay ratios

Whole copy numbers were generated by rounding the ratios obtained, from either qPCR or PRT, to the nearest integer, using conventional cut offs. Where the samples were tightly clustered around individual copy numbers with no overlap, such as with LTR (**Fig. 1e, f**) estimating the whole copy number is straightforward. However, this is rarely the case (**Fig. 1**). The *CCL4L1* assays have distinct clusters (**Fig. 1c, d**), but these are shifted left toward lower copy number, indicating a lower PCR efficiency for *CCL4L1* gene than for *CCL4*. For both the *CCL3L1* PRT assay and the qPCR assay the data are dispersed with little distinction between copy numbers. Hence, as the data did not cluster around integer copy numbers (**Fig. 1**) we also adopted k-means clustering to assign whole copy number. Cases and controls were scored separately using the "cluster k-means" algorithm in STATA. The number of clusters used equated to the number of distinct peaks in the distribution of assay ratios. Different numbers of means were examined by increasing the number of means up to as many as the saturated model (equates to the full range of whole numbers obtained by rounding). However, increasing the number of means had the effect of splitting single clusters into multiple clusters and in some instances, data from two different clusters were put into the same group. The saturated model, as expected, gave the worst grouping. Having examined the distributions, four means were used to score the PRT data at *CCL4L1*, the LTR and *CCL3L1* in cases whereas just three were used for *CCL3L1* in controls, which equates to the number of clusters obtained for each assay excluding those called as zero. We verified that the six duplicated samples with known copy number in the PRT experiment were consistent across plates and duplicates.

Hardy-Weinberg Equilibrium

An estimation maximisation (EM) approach was used to generate phased genotypes from total copy number in order to check for deviations from Hardy-Weinberg equilibrium. The observed and expected copy numbers were used to calculate the Pearson's χ^2 statistic to test for significant deviations from HWE.

Copy number genotypes, *i.e.* phasing of copy number into number of copies of *CCL3L1* on each chromosome, was done by using the observed copy number frequency as an estimate of the prior probabilities, $P_{i,}$ of having a specific `allele', *i*. By allele we mean copies per chromosome (CPC) of *CCL3L1* or *CCL4L1* or the LTR. These probabilities were used to calculate posterior probabilities of the genotype combinations assigned to individuals. So if an individual had three copies of *CCL3L1*, then they had probability $2P_0.P_3$ of having genotype 0/3 (*i.e.* 0 copies on one chromosome and three copies on the other chromosome) under HWE and $2P_1.P_2$ of having genotype 1/2 under HWE. The E step of the EM algorithm is to estimate the posterior probabilities of the genotypes, by summing the genotype probabilities together for each individual and scaling them back to unity. So for example, if a subject had three total copies of *CCL3L1*, they would have a posterior probability of $P_0P_3 / (P_0P_3 + P_1P_2)$ of having genotype 0/3 and a posterior probability $P_1P_2/(P_0P_3 + P_1P_2)$ of having genotype 1/2. The M (maximisation) step of the EM algorithm is to maximise the fit of the CPC frequencies to total copy number, which is achieved by regenerating the prior probabilities. Summing the posterior over every instance of the CPC copy number and scaling to unity across all instances of the copy number generates the priors. The log likelihood is calculated by taking logs of the un-scaled posteriors and summing over all individuals. The E and M steps were repeated until no further increase in the log likelihood could be achieved. The Pearson's χ^2 test statistic is on $G - A$ degrees of freedom where G is the number of genotypes and A is the number of alleles.

Association tests

With a sample size of 4,000 cases and 4,000 controls, we have greater than 80% power to find effect sizes as low as OR=1.2 at $\alpha = 1\text{x}10^{-5}$.

Association with T1D was tested using a logistic regression model with disease status as the outcome variable and total copy number as the dependent variable. Total copy number was coded either as a factor or separately as a continuous variable without having rounded to the nearest integer. Geographical region was included as a confounder in the logistic model. For the qPCR experiment, as cases and controls were arrayed on different plates, it was necessary to cluster by plate when testing for association with T1D.

Supplementary Results

Duplication of the 17q12 CNV

The PRT assay depends on consistency between the number of copies of *CCL3L1* and *CCL4L1* and the LTR, *i.e.* both *CCL3L1* and *CCL4L1* and the region between them, are copied in their entirety, so individuals will have the same number of copies of *CCL3L1*, *CCL4L1* and the LTR. Indeed this was found to be the case for the majority of our data. We found that 229 (6.4%) cases and 13 (0.3%) controls had one less copy of *CCL3L1* than *CCL4L1* while 31 cases and 50 controls had one or more additional copies of *CCL3L1*. The consistency between the LTR and *CCL4L1* was greatest, with 115 (1.4%) samples having one or more additional copies of *CCL4L1* than the LTR and 46 (0.6%) samples having one or more additional copies of the LTR. Conversely, 301 (4.1%) samples had additional copies of the LTR compared to *CCL3L1* and 82 (1.1%) samples had extra copies of *CCL3L1* than the LTR. Therefore, for the majority of data there were a consistent number of copies of each of *CCL3L1*, *CCL4L1* and the LTR.

Association of CCL3L1 with T1D using non-integer copy number

Expression of *CCL3L1* has been reported to correlate with copy number of *CCL3L1* in a dose dependent fashion [5]. We have assumed there is a correlation between integer copy number of *CCL3L1* and production of the protein CCL3L1. Hence, we initially analysed integer copy numbers of *CCL3L1* for association with T1D. Nevertheless, to remove the need to assign integer copy number, we also analysed *CCL3L1* assay ratios as a continuous trait. *CCL3L1*, from the qPCR assay, showed association with T1D when analysed as a continuous trait (P=5x10^{-5}). However, given that it deviated from HWE in controls, this was a false association and again illustrated the importance of checking for deviation from HWE and related genotyping errors. The LTR assay ratios, which were in HWE in controls, when analysed as a continuous trait were not associated with T1D ($P = 0.55$).

Deviation from HWE in small sample sets

We wished to check whether deviations from HWE could be detected in sample sets of a similar size to those used in other *CCL3L1* publications. Using our qPCR data in controls, we randomly selected 1,100 sample sets of an average size of 500, and tested each of them for deviations from HWE. (The full dataset significantly deviated from HWE, **Supplementary Table 3**.) Thirty one percent of these sample sets deviated from HWE at $P \leq 0.01$, implying that, with a sample size of 500, there is a 70% chance that deviations from HWE would not be detected.

PRT in African Yoruban samples

We also genotyped 95 African Yoruban samples in duplicate using the LTR PRT assay to compare with Gonzalez *et al.*[4] using the methods described above. Our data were highly reproducible (correlation coefficient > 0.99; **Supplementary Figure 2**). The CNVs obtained were between two and eight (**Supplementary Table 4**) with a mean copy number of 4.3 and a median copy number of 4.0. This is lower than the mean copy number reported by Gonzalez, which was six. We cannot comment on the performance of the assay in the Yoruban samples with respect to distribution of copy number around integer values due to insufficient numbers of samples.

Supplementary Discussion

Other studies have observed (using different technologies) small differences in sample chemistry can produce false positive associations[1,6,7]. These include, for qPCR, reports of amplicon efficiency having dependence on sample chemistry[6]. In our study this difference in DNA chemistry manifested as a differential shift in copy number in controls compared to cases using both qPCR and the *CCL3L1* PRT assay. In Gonzalez *et al.* the European-American HIV$^+$ cases and HIV$^-$ controls came from different sources[4]. Therefore, one explanation for the apparent strong disease protective effect of copy numbers higher than the population median they report, may be an artefact of a shift in distribution caused by chemical differences in DNA templates between cases and controls, similar to the difference observed in our data (**Figure 1**, **Supplementary Tables 1, 3**). This shift could have been confounded by the inclusion by Gonzalez *et al.* of a group of mixed European-Americans and Hispanic Americans, with Hispanic Americans showing higher *CCL3L1* copy numbers[4]. These effects together may account for the lack of replication of the associations reported by Gonzalez *et al.*[8].

The copy number of *CCL3L1* using qPCR is increased (right shifted) compared to the PRT LTR copy numbers. There is a known partial copy of *CCL3L1*, an untranslated pseudogene that does not include exon 1 or intron 1,[9] designated *CCL3L2*, which the Gonzalez *et al.*[4] qPCR assay cannot distinguish from *CCL3L1*. The PRT *CCL3L1* assay was specifically designed in intron 1 to avoid measuring this pseudogene[3] (**Supplementary Figure 1)**. Colobran *et al.*[10] and others have shown, in their qPCR assay, that the number of copies of *CCL3L1* does not necessarily equal the number of copies of *CCL4L1*[5,10]. Both the discrepancies observed in *CCL3L1* copy number counts between qPCR and the PRT *CCL3L1* assay, and the discrepancies between *CCL3L1* and *CCL4L1* copy number, may be due to the qPCR primers and probe binding to both exon 3 of *CCL3L1* and the pseudogene.

The PRT data did not cluster as well for *CCL3L1* as for either *CCL4L1* or the LTR. The poor clustering observed for *CCL3L1* on both assay formats is the result of inconsistent variation in PCR rates between samples. This unpredictable PCR rate may be the result of impurities in the DNA sample that interfere with *CCL3L1* PCR efficiency to varying extents in each sample, in particular, proteins bound to the *CCL3L1* sequence may inhibit the PCR reaction. Another possible factor is that loss of primer specificity may affect *CCL3L1* particularly, as one partial pseudogene is known and there may be other partial *CCL3L1* sequences elsewhere. The case and control DNA that we used were prepared in different laboratories using different (although similar) methods (involving a chloroform extraction step but not phenol). Therefore, we hypothesize that the differences in the composition between the case DNA and the control DNA differentially affect the efficiency of the assays in the two collections and led to an apparent false positive disease association. That the LTR copy number shows good clustering, and no difference in distribution between T1D cases and controls, suggests that the proposed protein contamination is not present for this sequence, and that the primers designed for this locus are highly specific and robust to variations in chemical composition of the DNA preparation.

Supplementary Table 1: Distribution of *CCL3L1* copy number in the 3,860 cases and 4,084 controls, and of *CCL4L1* copy number in 4,041 cases and 4,318 controls that were successfully genotyped using the *CCL3L1* PRT assay.

Copy #	*CCL3L1*						*CCL4L1*					
	Observed (Expected)				OR [95% CI]		observed (expected)				OR[95% CI]	
	Cases		Controls				Cases		Controls			
	Rounding	K-means	Rounding	K-means	Rounded	K-means	Rounding	k-means	Rounding	k-means	Rounding	k-means
0	71 (194.6)	71 (64.5)	74 (181.8)	74 (76.0)	1.13 [0.81-1.59]	0.97 [0.69-1.36]	78 (86.4)	78 (67.7)	79 (83.1)	79 (80.2)	1.13 [0.81-1.56]	1.05 [0.76-1.45]
1	1513 (1260.8)	761 (772.8)	1535 (1317.7)	828 (824.5)	1.16 [1.06-1.28]	0.93 [0.82-1.04]	962 (945.0)	764 (785.3)	995 (986.8)	876 (874.0)	1.10 [0.99-1.23]	0.92 [0.82-1.04]
2	1988 (2121.5)	2380 (2411.7)	2318 (2428.1)	2380 (2372.3)	1.00 [ref]	1.00 [ref]	2636 (2645.7)	2413 (2397.1)	2971 (2974.9)	2530 (2521.7)	1.00 [ref]	1.00[ref]
3	267 (261.9)	648 (575.3)	148 (146.9)	717 (737.1)	2.13 [1.72-2.64]	0.91 [0.81-1.03]	348 (343.7)	688 (702.0)	265 (265.6)	746 (765.8)	1.47 [1.24-1.75]	0.98 [0.87-1.11]
4	20 (20.4)	0	7 (9.3)	85 (72.2)	2.93 [1.23-6.98]		14 (19.7)	98 (84.1)	8 (7.6)	87 (74.2)	1.77 [0.73-4.28]	1.14 [0.84-1.53]
5	1 (0.8)	0	2 (0.2)	0	0.51 [0.04-5.73]		3 (0.6)	0	0 (0.7)	0		
6	0	0	0	0			0	0	0 (0)	0		
P_{HWE}	$1x10^{-30}$	0.0012	$9x10^{-27}$	0.088			0.0011	0.028	0.5435	0.0952		
P_{T1D}					$7.8x10^{-11}$	0.3686					0.0002	0.5688

OR = Odds Ratio. CI = Confidence Interval. P_{HWE} = Hardy-Weinberg Equilibrium *P*-value. P_{T1D} = *P*-value for association with type 1 diabetes.

Supplementary Table 2: Distribution of the chr17q12 LTR copy number in 4,044 cases and 4,266 controls successfully genotyped using the LTR PRT assay. Also given is the distribution of copy number averaged across the three PRT assays *CCL3L1*, *CCL4L1* and the LTR in 4,106 cases and 4,351 controls.

Copy #	*LTR*						Average across *CCL3L1*, *CCL4L1*, *LTR*					
	Observed (Expected)				OR [95% CI]		Observed (Expected)				OR[95% CI]	
	Cases		Controls				Cases		Controls			
	Rounding	K-means	Rounding	K-means	Rounded	K-means	Rounding	k-means	Rounding	k-means	Rounding	k-means
0	75 (62.2)	75 (62.7)	78 (73.9)	78 (75.6)	1.01 [0.73-1.40]	1.01 [0.73-1.41]	76 (67.0)	76 (64.0)	78 (75.1)	78 (76.5)	1.03 [0.74-1.43]	1.02 [0.74-1.42]
1	726 (752.3)	727 (752.6)	829 (837.4)	839 (844.5)	0.92 [0.82-1.03]	0.91 [0.81-1.02]	800 (818.3)	744 (768.9)	878 (884.7)	853 (856.6)	0.97 [0.86-1.08]	0.91 [0.81-1.02]
2	2408 (2393.3)	2404 (2382.5)	2510 (2505.0)	2510 (2496.8)	1.00 [ref]	1.00 [ref]	2602 (2592.0)	2454 (2432.8)	2725 (2726.3)	2554 (2539.9)	1.00[ref]	1.00 [ref]
3	732 (733.7)	735 (754.7)	756 (757.0)	752 (773.3)	1.03 [0.91-1.16]	1.04 [0.92-1.17]	574 (575.2)	735 (754.1)	629 (617.9)	778 (802.2)	0.97 [0.85-1.10]	1.00 [0.89-1.13]
4	93 (92.3)	103 (87.3)	83 (82.6)	87 (74.0)	1.17 [0.86-1.60]	1.22 [0.91-1.64]	52 (51.3)	97 (82.6)	38 (46.3)	88 (74.4)	1.43 [0.93-2.19]	1.12 [0.83-1.51]
5	9 (9.4)	0	9 (9.1)	0	1.04 [0.41-2.65]		2 (2.1)	0	3 (1.2)	0	0.51 [0.09-3.11]	
6	1 (0.8)	0	1 (0.9)	0	0.76 [0.05-12.40]		0	0	0	0		
P_{HWE}	0.1562	0.0092	0.8242	0.0807			0.4331	0.0126	0.0967	0.0678		
P_{T1D}					0.6946	0.2394					0.5588	0.4969

OR = Odds Ratio. CI = Confidence Interval. P_{HWE} = Hardy-Weinberg Equilibrium *P*-value. P_{T1D} = *P*-value for association with type 1 diabetes.

Supplementary Table 3: Distribution of *CCL3L1* copy number in the 3,362 cases and 3,983 controls that were successfully genotyped on both the original and replicate plates using qPCR.

Copies	Rounded			k-means clustering		
	Observed (expected)		OR [95%CI]	Observed (expected)		OR [95%CI]
	Cases	Controls		Cases	Controls	
0	53 (56.9)	29 (68.3)	2.17 [1.37-3.45]	53 (56.1)	29 (80.7)	1.83 [1.14-2.92]
1	598 (590.9)	750 (668.6)	0.85 [0.73-0.98]	611 (603.3)	778 (668.9)	0.70 [0.60-0.81]
2	1610 (1610.5)	1729 (1771.8)	1.00 [ref]	1684 (1693.1)	1517 (1574.8)	1.00 [ref]
3	467 (461.5)	746 (739.8)	0.67 [0.51-0.89]	413 (442.1)	803 (847.8)	0.46 [0.33-0.64]
4	399 (400.3)	453 (451.5)	0.94 [0.79-1.11]	430 (377.8)	486 (441.3)	0.79 [0.66-0.94]
5	158 (169.3)	173 (175.8)	0.98 [0.69-1.39]	132 (148.4)	281 (273.8)	0.44 [0.31-0.62]
6	45 (51.8)	66 (80.9)	0.81 [0.51-1.29]	41 (30.1)	89 (69.1)	0.46 [0.28-0.76]
7	25 (15.9)	23 (18.8)	1.26 [0.65-2.45]			
8	6 (4.2)	12 (6.0)	0.54 [0.21-1.44]			
9	1 (0.7)	1 (1.2)	0.77 [0.04-15.9]			
10	0 (0.0)	1 (0.3)	1.59 [0.09-27.8]			
P_{HWE}	0.1454	$1\text{x}10^{-8}$		0.0005	$5\text{x}10^{-15}$	
P_{T1D}			0.0002			$7\text{x}10^{-9}$

OR = Odds Ratio. CI = Confidence Interval. P_{HWE} = Hardy-Weinberg Equilibrium *P*-value. P_{T1D} = *P*-value for association with type 1 diabetes.

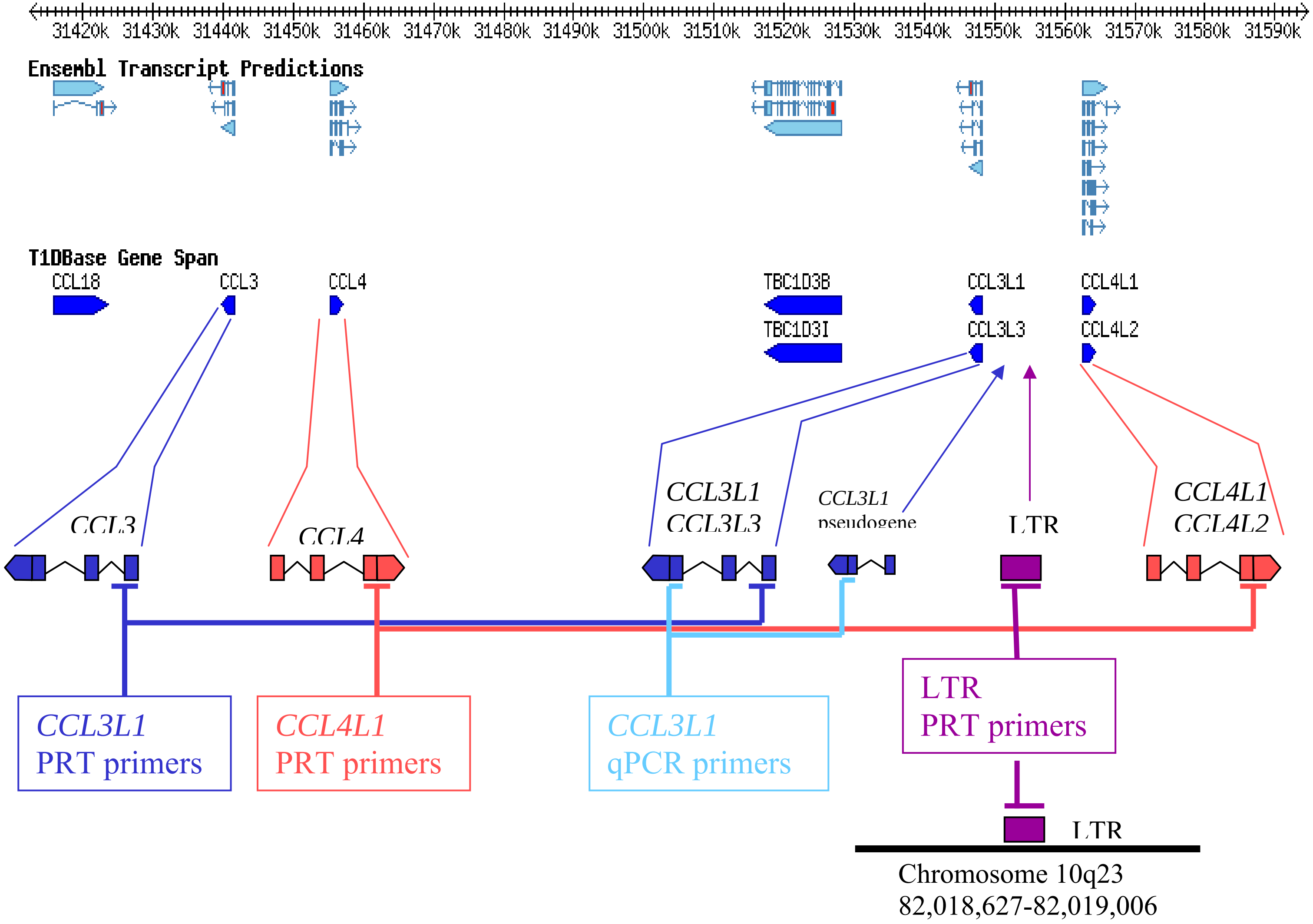
Supplementary Table 4: Distribution of the chromosome 17q12 LTR copy number in 91 Yoruban samples successfully genotyped in duplicate using the LTR PRT assay and then rounding to the nearest integer.

Copies	N (%)
< 2	0 (0)
2	10 (11.0)
3	17 (18.7)
4	26 (28.6)
5	17 (18.7)
6	17 (18.7)
7	3 (3.3)
8	1 (1.1)

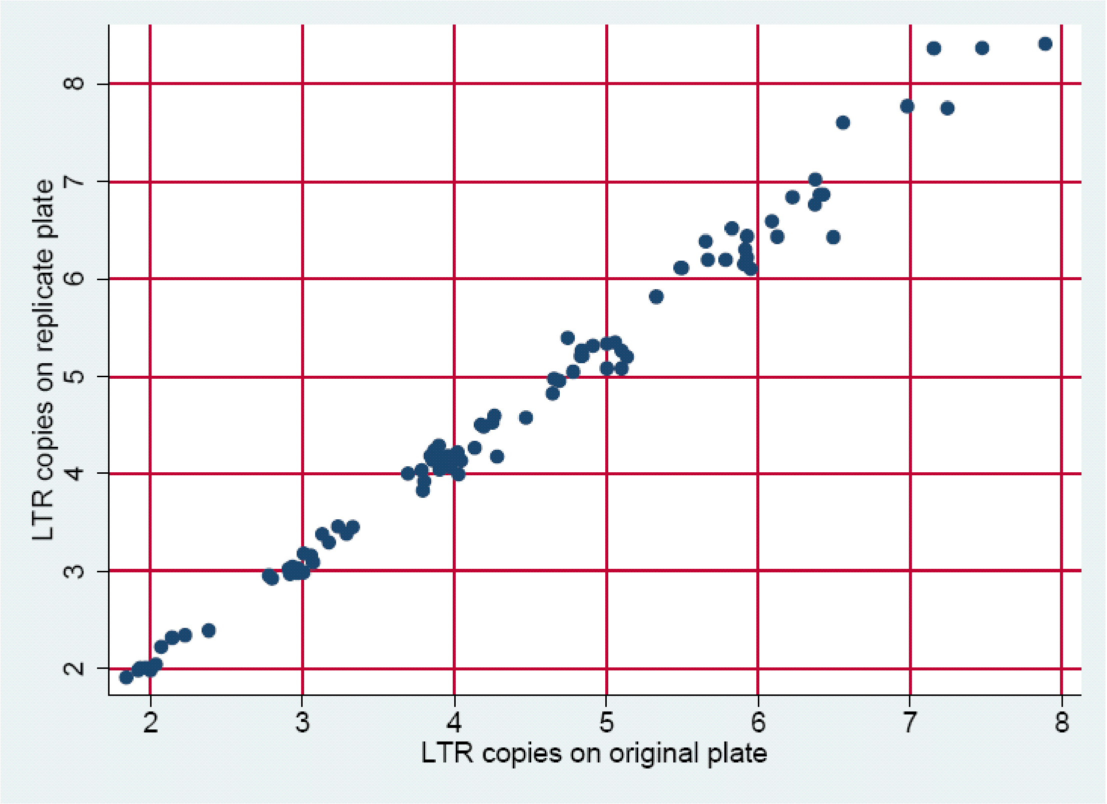
Supplementary Table 5: Distribution of copy numbers obtained by the LTR PRT assay in samples that failed on either the "original" or "duplicate" plates, showing that there is no bias in copy number frequency among PCR failures.

Copies	Frequency of "originals"	Failed on "original" but passed on "duplicate"	Frequency of "duplicates"	Failed on "duplicate" but passed on "original"
0	159 (1.8)	7 (1.1)	146 (1.8)	13 (2.5)
1	3,346 (38.7)	258 (38.9)	3,095 (38.1)	200 (38.5)
2	4,598 (53.2)	339 (51.1)	4,367 (53.8)	269 (51.8)
3	500 (5.8)	56 (8.4)	479 (5.9)	34 (6.6)
4	35 (0.4)	4 (0.6)	33 (0.4)	3 (0.6)
5	6 (0.1)		5 (0.1)	

Supplementary Figure 1: The copy number variant region on chromosome 17q12, available from www.T1DBase.org (from NCBI build 130). Predicted primer binding sites for the PRT primers and the qPCR primers and probe are also given. *CCL3L3* and *CCL4L2* are variants of *CCL3L1* and *CCL4L1* respectively that are defined by single nucleotide polymorphisms, neither the PRT nor the qPCR assays can distinguish these variants. Thus the *CCL3L1* PRT primers recognise *CCL3*, *CCL3L1* and *CCL3L3* and the *CCL4L1* PRT primers recognise *CCL4*, *CCL4L1* and *CCL4L2*. The *CCL3L1* pseudogene is a truncated version of *CCL3L1* lacking exon 1, as the qPCR primers and probes bind to exon 3 they recognise *CCL3L1*, *CCL3L3* and the pseudogene.



Supplementary Figure 2: Comparison of copy number obtained for the 95 Yoruban samples genotyped using the LTR PRT assay on original and replicate plates.



References

1. Clayton, D.G., *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet* 37, 1243-1246 (2005).
2. Armour, J.A., *et al.* Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. *Nucleic Acids Res* 35, e19 (2007).
3. Walker, S., Janyakhantikul, S. & Armour, J.A. Multiplex Paralogue Ratio Tests for accurate measurement of multiallelic CNVs. *Genomics* (2008).
4. Gonzalez, E., *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* 307, 1434-1440 (2005).
5. Townson, J.R., Barcellos, L.F. & Nibbs, R.J. Gene copy number regulates the production of the human chemokine CCL3-L1. *Eur J Immunol* 32, 3016-3026 (2002).
6. Karlen, Y., McNair, A., Perseguers, S., Mazza, C. & Mermod, N. Statistical significance of quantitative PCR. *BMC Bioinformatics* 8, 131 (2007).
7. WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661-678 (2007).
8. Shao, W., *et al.* CCL3L1 and CCL4L1: variable gene copy number in adolescents with and without human immunodeficiency virus type 1 (HIV-1) infection. *Genes Immun* 8, 224-231 (2007).
9. Modi, W.S. CCL3L1 and CCL4L1 chemokine genes are located in a segmental duplication at chromosome 17q12. *Genomics* 83, 735-738 (2004).
10. Colobran, R., *et al.* Population structure in copy number variation and SNPs in the CCL4L chemokine gene. *Genes Immun* 9, 279-288 (2008).