IMMEDIATE COMMUNICATION

Visual analysis of geocoded twin data puts nature and nurture on the map

OSP Davis[1,2,3], CMA Haworth[1], CM Lewis[1] and R Plomin[1]

Twin studies allow us to estimate the relative contributions of nature and nurture to human phenotypes by comparing the resemblance of identical and fraternal twins. Variation in complex traits is a balance of genetic and environmental influences; these influences are typically estimated at a population level. However, what if the balance of nature and nurture varies depending on where we grow up? Here we use statistical and visual analysis of geocoded data from over 6700 families to show that genetic and environmental contributions to 45 childhood cognitive and behavioral phenotypes vary geographically in the United Kingdom. This has implications for detecting environmental exposures that may interact with the genetic influences on complex traits, and for the statistical power of samples recruited for genetic association studies. More broadly, our experience demonstrates the potential for collaborative exploratory visualization to act as a *lingua franca* for large-scale interdisciplinary research.

INTRODUCTION

Twin and family studies are an important counterpart to the genomic revolution that has taken place since the sequencing of the human genome.[1,2] Although molecular genetic techniques, such as genome-wide association and sequencing, have the advantage of allowing us to identify individual genetic variants that are important for population variation in traits and diseases, twin and family studies have the advantage of taking into account all DNA variation throughout the genome and the population, simply by using what we know about genetic relatedness within twin pairs of different zygosity.[3,4]

One obvious contribution of twin and family studies to genetics in the postgenomic era has been in setting the benchmark for studies that aim to identify genetic variants that have a role in the inheritance of complex traits.[2] The gap between the population variance accounted for by the current catalog of variants and the variance we estimate to be accounted for by genetic effects has come to be known as 'missing heritability'. The search for the sources of missing heritability has spawned scientific innovation and collaboration on a global scale.[5]

The second great advantage of twin and family studies is that they give us insight into the other side of the etiology of complex traits and disorders that is completely invisible to DNA microarrays and next-generation sequencing platforms: the action of the environment on variation at the population level. In the midst of the genomic revolution, it is easy to forget that even under the optimistic premise that we will eventually identify every genetic variant that influences a complex trait, we will still only know half the story of its origin, because identifying the influential environments that account for the remaining variance is just as important. This array of physical and psychosocial environmental exposures has been characterized as an 'exposome' that parallels the genome in its influence on complex traits.[6] As with genetic variation, twin and family studies are agnostic to the many forms these environments may take. So, although we have not yet identified all the important elements of the exposome (there are 'missing environments' as well as missing heritability), twin studies still allow us to explore the mass influence of environmental variation on the phenotype.

Similar to early forays into molecular genetics, early twin studies were often victims of small sample sizes and limited technology. However, the modern twin studies of recent decades often number in the thousands, or tens of thousands, of participants, and take advantage of advances in maximum likelihood structural equation modeling to fit sophisticated etiological models. These studies have proved robust to methodological challenges[2] and produce results that replicate consistently to identify important aspects of the joint action of genes and environments. These analyses allow us to carve nature and nurture at the joints, suggesting targeted hypotheses for our studies of specific genetic and environmental variation.

One aspect of these large, population-based epidemiological samples that has remained unexplored is how geographical location can affect the influence of nature and nurture on a phenotype. We aimed to address this question using geocoded data on 45 phenotypes collected at age 12 from 6759 twin pairs participating in the Twins Early Development Study (TEDS).[7] However, making sense of the genetic and environmental etiology of childhood traits and disorders is a complex process that requires input from experts in a wide range of fields, so we sought an approach that could capitalize on this distributed expertise. Here we describe a novel twin modeling approach that incorporates spatial information, and the design of the *spACE* interactive visual analysis tool that allowed us to collaboratively

[1]King's College London, MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, London, UK; [2]The Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK and [3]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK. Correspondence: Dr OSP Davis, King's College London, MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, De Crespigny Park, London SE5 8AF, UK.
E-mail: Oliver.Davis@kcl.ac.uk

explore the geocoded twin data, with contributions from geneticists, psychologists, statisticians, clinicians, geographers and teachers.

MATERIALS AND METHODS

Twins Early Development Study

The UK Office for National Statistics contacted parents of all twins born in England and Wales between 1994 and 1996, and invited them to take part in a longitudinal investigation of genetic and environmental influences on behavior and cognition. The vast majority (over 12 000 families) agreed to take part and have been followed over the first 16 years of life. To date, they remain representative of the UK population through comparison with census data.[7] Most have remained in England and Wales, although a few have migrated to Scotland or overseas. In keeping with the UK population, most (96%) identify themselves as white British with English as their first language; for this analysis, we included only these families, to avoid greater genetic heterogeneity in cities biasing our results. Ethical approval was provided by the Institute of Psychiatry Ethics Committee of King's College London.

When the twins were 12 years of age, we carried out our broadest survey to date, using Web-based testing,[8] parent questionnaires and teacher reports to assess a wide range of cognitive abilities, behavioral (and other) traits, environments and academic achievement in 6759 reared-together twin pairs. The 45 phenotypes included in this study are described in Table 1.

Spatial statistics and visual analysis

All 6759 pairs of twins were assigned geographical coordinates using the UK National Postcode Database. As the UK postcodes are generally unique to a street or a small group of addresses, this gave us an accurate location to a sub-neighborhood scale. We used maximum likelihood structural equation modeling of the twin data in OpenMx[9] to calculate genetic and environmental contributions to a range of childhood phenotypes assessed at 12 years of age (see Table 1), for a series of target locations across the United Kingdom. The locations were chosen to represent the local density of twin data, to provide a visual cue to the data available for the estimation of variance components in different areas. However, to preserve anonymity, they do not correspond to the locations of individual twin pairs. Every family contributed to each calculation, but with a weight assigned according to spatial proximity to the target location (Figure 1). Each twin pair's contribution to the analysis was weighted by the inverse of their distance from the point of estimation:

$$w_i(x) = \frac{1}{d(x, x_i)^p}$$

where x represents the point of estimation, x_i represents the location of a twin pair, d is the Euclidean distance between x and x_i, and p is the power parameter (0.5 for these analyses). We applied the weights by calculating weighted covariance matrices for monozygotic and dizygotic twin pairs, and using these as an input for the structural equation twin model, iterating over the target locations. Supplementary Figures 1 and 2 describe a series of simulations we carried out to test this approach using an artificial data set with known parameters.

Following principles of human perception and interaction design,[10] we developed a purpose-built visual analysis tool programmed in the Processing visualization language to display and explore the output, as described in Figure 2. This revealed patterns of geographical variation in genetic and environmental contributions to the childhood phenotypes. The *spACE* visualization tool is available as Supplementary Software, loaded with the 45 TEDS phenotypes, from http://sgdp.iop.kcl.ac.uk/davis/teds/geocoding/, where there is also a video demonstrating the

Table 1. Description of the measures included in the analysis

Phenotype name	Description	References
Language	Composite of three language tests	8
Reading	Composite of four reading tests	8
Verbal ability	Composite of two verbal tests	8
Nonverbal ability	Composite of two nonverbal tests	8
IQ	Composite of two verbal tests and two nonverbal tests	8
Mathematics	Composite of three mathematics tests	8
School English	Teacher assessment of English attainment, with reference to the Key Stages of the UK National Curriculum	http://curriculum.qcda.gov.uk/
School Mathematics	Teacher assessment of mathematics attainment, with reference to the Key Stages of the UK National Curriculum	http://curriculum.qcda.gov.uk/
School Science	Teacher assessment of science attainment, with reference to the Key Stages of the UK National Curriculum	http://curriculum.qcda.gov.uk/
School achievement	Teacher assessment of educational attainment, with reference to the Key Stages of the UK National Curriculum (composite of the three core subjects)	http://curriculum.qcda.gov.uk/
Parent ASD social	Social subscale of the Childhood Asperger Syndrome Test (parent-rated)	27,28
Teacher ASD social	Social subscale of the Childhood Asperger Syndrome Test (teacher-rated)	27,28
Parent ASD nonsocial	Nonsocial subscale of the Childhood Asperger Syndrome Test (parent-rated)	27,28
Teacher ASD nonsocial	Nonsocial subscale of the Childhood Asperger Syndrome Test (teacher-rated)	27,28
Parent ASD comms	Communication subscale of the Childhood Asperger Syndrome Test (parent-rated)	27,28
Teacher ASD comms	Communication subscale of the Childhood Asperger Syndrome Test (teacher-rated)	27,28
Parent ASD total	Childhood Asperger Syndrome Test composite (parent-rated)	27,28
Teacher ASD total	Childhood Asperger Syndrome Test composite (teacher-rated)	27,28

Table 1 (*Continued*)

Phenotype name	Description	References
ADHD hyperactivity	Hyperactivity subscale of the Conners' Parent Rating Scale	29
ADHD inattention	Inattention subscale of the Conners' Parent Rating Scale	29
ADHD total	Conners' Parent Rating Scale composite	29
Moods and feelings	Moods and Feelings Questionnaire (parent-rated)	30
Parent prosocial	Prosocial subscale of the Strengths and Difficulties Questionnaire (parent-rated)	31,32
Teacher prosocial	Prosocial subscale of the Strengths and Difficulties Questionnaire (teacher-rated)	31,32
Parent hyperactivity	Hyperactivity subscale of the Strengths and Difficulties Questionnaire (parent-rated)	31,32
Teacher hyperactivity	Hyperactivity subscale of the Strengths and Difficulties Questionnaire (teacher-rated)	31,32
Parent conduct	Conduct subscale of the Strengths and Difficulties Questionnaire (parent-rated)	31,32
Teacher conduct	Conduct subscale of the Strengths and Difficulties Questionnaire (teacher-rated)	31,32
Parent peers	Peers subscale of the Strengths and Difficulties Questionnaire (parent-rated)	31,32
Teacher peers	Peers subscale of the Strengths and Difficulties Questionnaire (teacher-rated)	31,32
Parent emotional	Emotional subscale of the Strengths and Difficulties Questionnaire (parent-rated)	31,32
Teacher emotional	Emotional subscale of the Strengths and Difficulties Questionnaire (teacher-rated)	31,32
Parent behavior	Composite of the problem behavior subscales from the Strengths and Difficulties Questionnaire (parent-rated)	31,32
Teacher behavior	Composite of the problem behavior subscales from the	31,32

Table 1 (*Continued*)

Phenotype name	Description	References
	Strengths and Difficulties Questionnaire (teacher-rated)	
Parent callous	Callous-unemotional subscale of the Antisocial Process Screening Device (parent-rated)	33,34
Teacher callous	Callous-unemotional subscale of the Antisocial Process Screening Device (teacher-rated)	33,34
Parent narcissism	Narcissism subscale of the Antisocial Process Screening Device (parent-rated)	33,34
Teacher narcissism	Narcissism subscale of the Antisocial Process Screening Device (teacher-rated)	33,34
Parent impulsivity	Impulsivity subscale of the Antisocial Process Screening Device (parent-rated)	33,34
Teacher impulsivity	Impulsivity subscale of the Antisocial Process Screening Device (teacher-rated)	33,34
Parent antisocial	Composite of the Antisocial Process Screening Device (parent-rated)	33,34
Teacher antisocial	Composite of the Antisocial Process Screening Device (teacher-rated)	33,34
Height	Height in meters	
Weight	Weight in kilograms	
BMI	Weight in kilograms, divided by height in meters squared	

Abbreviations: ASD, autism spectrum disorder; ADHD, attention deficit hyperactivity disorder; BMI, body mass index.

visualization. The software download includes an OpenMx script for the R statistical computing environment that demonstrates our approach to geographically sensitive twin models.

Supplementary statistical models

Fitting further structural equation models can formally test these patterns for statistical significance. In the example below, we explore the relationship between income inequality and classroom behavior. To do this, we used the continuous moderator twin model that allows the contribution of genetic and environmental variance components to vary as a function of a measured environment[11] (Supplementary Figure 5). Testing the significance of the moderation term allows us to establish moderation of the genetic and environmental effects by the measured environment. Simulations suggest that our sample for this analysis (the 5073 pairs of twins with matching phenotype data) gives us 80% power to detect a moderation of the E-term of the size we observed. For more discussion of the continuous moderator model, see Hanscombe *et al.*[12] In our example, we test for the moderation of the variance components by local variance in household
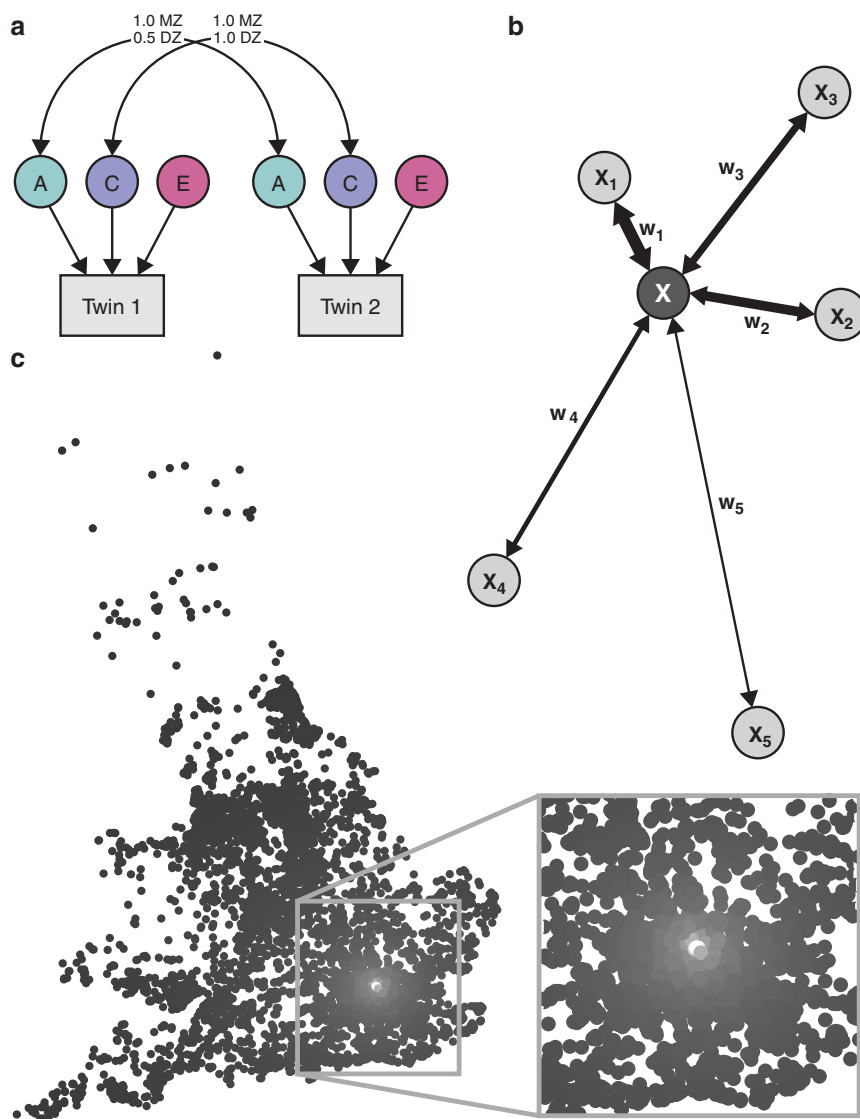
Figure 1. Calculation of genetic and environmental influences at a geographical location. (**a**) The structural equation model is based on a standard univariate twin model that partitions the phenotypic variance into additive genetic influences (A), shared (common) environmental influences (C) that make children in the same family similar to each other, and non-shared environmental influences (E) that do not contribute to similarity within families. We are able to do this because A influences are 100% shared between monozygotic (MZ) twins, whereas they are shared on average 50% between dizygotic (DZ) twins. In contrast, the C component is 100% shared by both MZ and DZ twins, and the E component is not shared at all. These components are calculated for a series of geographical locations. (**b**) For each geographical location (x) the same model is fitted, with each twin pair's (x_i) contribution to the analysis weighted (W_i) according to the inverse of their Euclidean distance from x, as described in the Materials and Methods. An OpenMx script implementing this is available as part of the spACE software download. (**c**) A gray scale demonstrates the relative contributions of twin pairs to an analysis conducted at the highlighted point. The lighter points near to the target location contribute more to the analysis, with influence falling off with distance from the target location. All participants contribute to the analysis at every location; it is only their relative weight that changes.

income, estimated by calculating the variance in household income from our population sample, weighted by geographical distance in the same way as our twin model.

RESULTS

The main outcome of this study is the collection of interactive maps that highlight genetic and environmental hotspots for a wide range of psychiatrically relevant childhood phenotypes, available for full exploration at http://sgdp.iop.kcl.ac.uk/davis/teds/geocoding/. There are many research questions that may be

addressed using visual analysis of these data, and we hope that looking at one example in detail here will encourage interest in the corresponding maps for components of the autism spectrum, attention-deficit hyperactivity disorder, mood, or cognitive abilities, such as reading, mathematics or general cognitive ability, for instance. A full list of the 45 phenotypes is provided in Table 1.

As an example, here we explore the relationship between income inequality, classroom behavior and educational outcomes in early adolescence. A notable pattern in our plots of the geographical distribution of genetic and environmental influences is a trend towards greater environmental influence on classroom
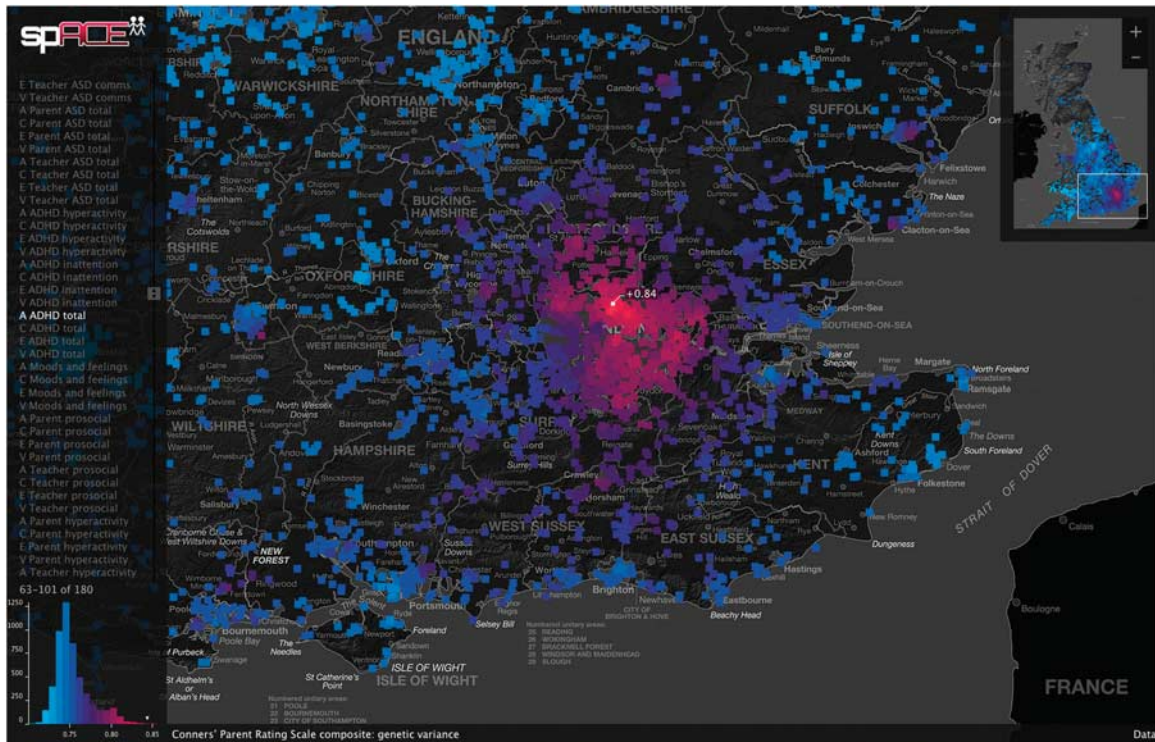
Figure 2. Visual analysis of geocoded twin data. We used the Processing visualization language to develop an interactive environment for exploring the geographical patterns of nature and nurture (available from http://sgdp.iop.kcl.ac.uk/davis/teds/geocoding/ and demonstrated in the video there). A divergent blue (low) to red (high) color palette indexes the variance attributable to genetic or environmental effects, with increased luminance at the extremes helping to emphasize areas that diverge from the national average. A small map of the whole United Kingdom provides an overview of the pattern, whereas the panning and zooming main display provides a closer view. Read-off of the exact values is provided on mouseover, and the value is linked to a color-coded histogram of data from the whole map to anchor the color scale. Maps of genetic and environmental influences on a range of childhood phenotypes are selected from a list to the left of the screen, and a more detailed description of the current map appears along the bottom of the screen. This visual approach has allowed researchers from a wide range of disciplines to collaborate in generating and investigating hypotheses about why these genetic and environmental hotspots occur, suggesting candidates for formal statistical testing.

behavior variables and academic achievement in London, compared with the rest of the UK. We hypothesized that the metropolitan area has a greater juxtaposition of extremely rich and extremely deprived neighborhoods than in any other part of the country, and that local variability in household income during childhood could be contributing to the increased environmental influence in the London area. Plotting the locally weighted variance in household income (an index of regional income inequality) side-by-side with the environmental influence on the total classroom behavior problems index, as in Figure 3, fits with this hypothesis, and Figure 4 shows one plotted against the other. Supplementary Figures 3 and 4 explore how likely it is that the pattern of results in Figure 3 occurred by chance. We observe the same pattern for teacher-rated narcissism and academic performance against National Curriculum targets, indicators that have previously been associated with income inequality on both national and international scales.[13] Of course, this correlation does not necessarily demonstrate causality, although plausible mechanisms linking these indicators have been suggested, both at the psychological level (with inequality leading to status anxiety, narcissistic self-promotion, deterioration of classroom behavior and poor academic achievement[14]), and at the neural level.[15,16] Using structural equation models, it is possible to fit environmental variables, such as local income inequality, as continuous moderators of the genetic and environmental parameters of the twin model.[11] Fitting these models confirms a

significant relationship between income inequality and the other measures (see Figure 4, and Supplementary Figures 5 and 6).

DISCUSSION

How should geographical variation in genetic and environmental effects be interpreted? In some ways, geographical variation in environmental effects seems more straightforward; finding an environmental hotspot simply suggests that environmental variation has more effect on the phenotype in that region. However, in our example, local variation in household income seemed to moderate the non-shared environmental effect rather than the shared environmental effect. Surely, local variation in household income is an environment shared by children growing up in the same family? One of the great strengths of the twin method is that it tells us at least as much about the environment as it does about genetics. An early and surprising finding from modern, well-powered twin cohorts was that the same environments are often experienced differently by children growing up in the same family.[17] Even environments that would intuitively be classed as shared environments end up making children in the same family different. This is particularly easy to imagine with an environment such as local variation in household income; in areas with high variation, it would be easier for children from the same family to make friends with peers from quite different backgrounds. In fact, it is usually not possible to classify particular
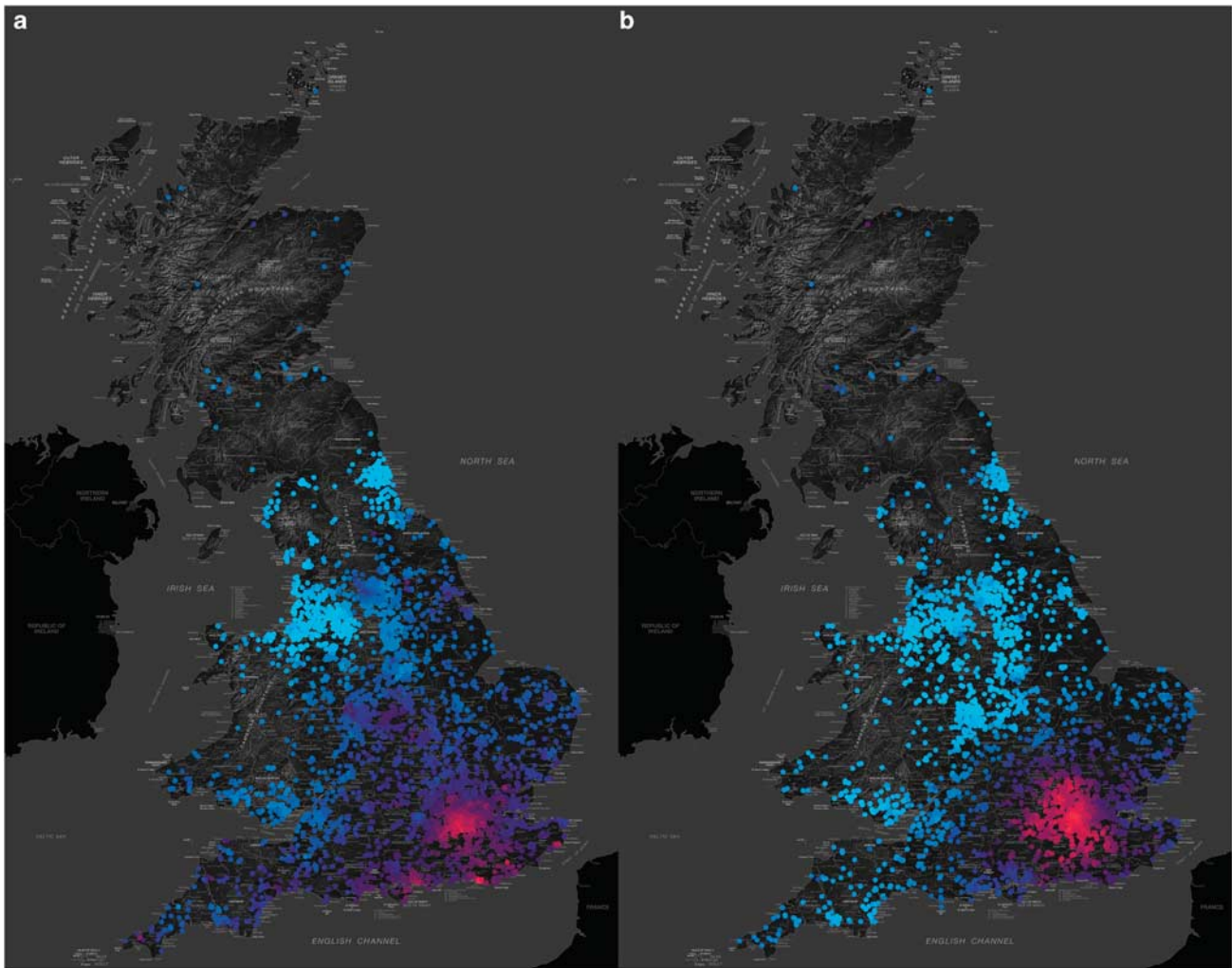
Figure 3. Classroom behavior problems and local variance in income. (**a**) Geographically weighted twin analysis of a composite measure of teacher-rated behavior problems suggests a greater contribution of environmental variance in London than in the rest of the United Kingdom. This map shows the distribution of the non-shared environmental (E) component of the twin model. The variance component varies from a high of 0.46 in London (red) to a low of 0.36 in the north east of the country (blue). (**b**) Our approach allows us to compare the distribution of candidate environments with the distribution of the environmental variance components. This map shows the locally weighted variance of household income, with greater income variance in London (red) compared with the rest of the United Kingdom (blue).

environments as either shared or non-shared, because it is likely that any one environment will have a mix of effects, some shared by members of the same family and some not.

Geographical differences in genetic effects seem more difficult to interpret, because it is unlikely that there are large genetic differences between people living in different regions of the United Kingdom. However, we should remember that the map shows not genetic differences, but differences in the effect that DNA variation has on a phenotype; we are looking at environmental moderation of genetic effects. Regions with large genetic effects are regions where the environment allows genetic variation to express itself through the phenotype. For example, imagine living next door to a field of wind-pollenated crops; in this situation, genetic variants that increase or decrease the risk of hayfever will lead to more phenotypic variation in the population than in an area where no pollen is present (where no one suffers from hayfever, irrespective of his or her genotype). In this way, presence of a strong genetic effect could in fact represent the presence of an environment that reveals relevant genetic variation (or the absence of a masking environment). We should note that in our model, the genetic and environmental parameters are not

constrained to add up to one at any one location, so it is possible for a region to be both a genetic and an environmental hotspot at the same time.

Clearly, like any technique, our approach has limitations. For example, one of the main purposes of this method is to nominate environments that interact with genetic and environmental influences on traits. However, it will only be able to identify environments that are geographically distributed on a large scale. Such environments may include social environments, such as district-level health, education provision or the psychosocial stress of urban living,[18] or physical environments, such as water or air pollution.[19] But many environments vary on much smaller scales; noise from an airport or main road may have a discernable influence, but the effect of noisy neighbors will not be apparent on a map of this resolution. In some cases it may help to follow up the visualization with further statistical models where a specific environment is measured at a fine-grained level. In other cases, these small-scale environments may go undetected by our approach. In addition, the twin method, though useful, has idiosyncrasies. For example, not all of the variance components are estimated with equal power; the shared environment is
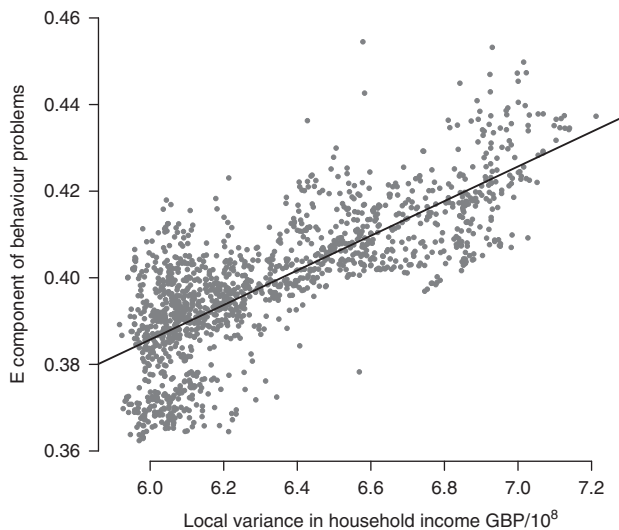
Figure 4. Environmental influence on teacher-rated behavior problems is related to income inequality. A scatter plot of environmental variance against locally weighted variance in household income confirms the relationship in Figure 3, plotted with the linear regression line ($y = 0.040x + 0.146$). This association can be formally tested using a twin model that allows the genetic and environmental variance components to vary as a function of a specific environment (11, Supplementary Figure 5). Fitting this model to the classroom behavior problems, narcissism and academic achievement variables reveals moderation of the non-shared environmental component by local variance in income (Supplementary Figure 6): for classroom behavior problems (difference in log likelihood = 10.3; degrees of freedom = 1; P-value = 0.0013), teacher-rated narcissism (9.22; 1; 0.0024), academic achievement for English (4.98; 1; 0.026), Mathematics (7.64; 1; 0.0057), Science (10.8; 1; 0.00099), and total academic achievement (12.3; 1; 0.00046).

particularly difficult to detect and the non-shared environmental term also encompasses measurement error (although this seems unlikely to vary geographically). More generally, as we saw earlier, the environmental terms are open to misinterpretation; it is sometimes remarked that the twin method does not take into account epigenetic variation. The truth is that the twin model does include epigenetic effects, but they (at least effects that are not inherited cross-generation) are included as environments, because the environment incorporates everything other than DNA sequence. In fact, this has fueled fascinating speculation that epigenetic variation may mediate long-term environmental effects on phenotypes.[20] Again, it seems that epigenetic factors are unlikely to vary geographically unless they are mediating some environmental effect, so this will probably not impact our findings. Finally, two characteristics of our sample deserve comment. The first is that some of our measures comprise rating scales completed by class teachers. Where twins share a class, they will likely share a rater. Where they do not share a class, they will not. However, in our sample we do not find that monozygotic twins share a classroom any more often than dizygotic twins, and our data suggest no systematic regional effect on sharing. The second characteristic is that TEDS is currently an adolescent sample, so the twins have had relatively little freedom to seek out geographical regions correlated with their genetic propensities. This means that we are limited in our ability to explore the effects of gene–environment correlation.[21] Tracking how individuals seek out new environments will be a fascinating topic for future waves of the study.

Our demonstration that the genetic and environmental contributions to complex traits vary depending on where we live has implications for several fields. For example, one

plausible explanation of the missing heritability encountered by genome-wide association studies has been gene-by-environment interactions. Our findings suggest a new mechanism for the identification of relevant environments, where exposure to the environment correlates with geographical distribution. Our findings also imply that when we conduct studies to identify variants associated with complex traits, the area where the sample is recruited will influence the power of our analysis to detect the variants. To take our earlier example, because the environment accounts for more of the variance in childhood behavior problems in London, this environmental variability will have a greater effect in masking the genetic associations in a sample recruited in London than it will in the rest of the United Kingdom. Researchers have invested a great deal of effort in identifying endophenotypes that are more heritable than their trait of interest to maximize their chances of finding genetic associations.[22,23] Our findings suggest that a similar principle applies to geographical location; as well as considering the logistics of sample collection in recruiting participants for association studies, researchers may also benefit from considering the variability of relevant environments in their catchment area. Recruiting in areas where the trait of interest is more heritable is likely to improve power to detect the effects of individual genetic variants. In addition, new analyses of genome-wide association data have been used to estimate the proportion of heritability captured by genotyping arrays.[24] Our approach is likely to translate directly to this type of data, reproducing our twin study results using molecular genetic information. This type of analysis will also translate to other forms of 'distance'. For example, it is easy to imagine applications in which time-of-travel is used instead of Euclidean distance, or even maps in which the axes represent conceptual dimensions rather than geographical ones.

Recently, there has been growing interest in the potential of visualization to take advantage of the broad information bandwidth of the human visual system in integrating large amounts of scientific data and spotting patterns amongst complexity.[25] One particularly interesting aspect of this, that we believe will gain further ground in coming years, is the trend towards integrating visualization into the analytic process, instead of approaching it as a way to effectively communicate the outcome of a completed study.[26] Alongside this, there is increasing recognition of how principles of visual perception, human–machine interaction and design can help us to construct visual analysis tools in a way that complements the idiosyncrasies of the way we perceive the world. Just as with statistical analysis, there are principles that must be followed to produce a valid result.[10] For example, one challenge we faced in designing our interactive map was choosing a color scale. We opted for a two-color scale that diverges from the mid-range, because this highlights the extremes of the distribution of values, picking out the high and low points that were most interesting in our analysis. The colors we chose were red and blue, avoiding red and green, because around 8% of men have a genetic mutation that makes it impossible for them to distinguish the two. Even so, color scales can be prone to optical illusions, such as simultaneous color contrast, in which the color we perceive is affected by the surrounding colors. To overcome this, we anchored the color scale, displaying numeric values on mouseover while indicating the point's position in the full distribution of values in a histogram to the left. Such challenges aside, new open-source initiatives such as the R project for statistical computing (http://www.r-project.org/) and the Processing visualization language (http://www.processing.org/) mean that it is now possible to construct a purpose-made custom visual analysis tool as part of the analytic process, an undertaking that previously would often have required too great an investment of time and resources to be practical.

Embedding visualization into our analysis has been crucial to arrive at these insights and connections. By bringing together experts from many different disciplines in the pursuit of a

common goal, it revealed patterns that merited further exploration. Beyond advancing our understanding of how nature and nurture interact on a national scale in the origins of childhood traits, we predict that collaborative visualization will play an increasingly important role in the scientific community's efforts to overcome the modern data deluge and arrive at an integrated understanding of complex systems.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

Author contributions

OSPD conceived and designed the study, performed the analysis and wrote the manuscript and software. RP and CMAH are Director and Deputy Director of TEDS. CML is Director of King's College London's Statistical Genetics Unit. All authors discussed the results, evaluated the software, and commented on the manuscript.

REFERENCES
1 Haworth CMA, Plomin R. Quantitative genetics in the era of molecular genetics: learning abilities and disabilities as an example. *J Am Acad Child Adolesc Psychiatry* 2010; **49**: 783–793.
2 Visscher PM, Hill WG, Wray NR. Heritability in the genomics era—concepts and misconceptions. *Nat Rev Genet* 2008; **9**: 255–266.
3 Boomsma D, Busjahn A, Peltonen L. Classical twin studies and beyond. *Nat Rev Genet* 2002; **3**: 872–882.
4 Martin N, Boomsma D, Machin G. A twin-pronged attack on complex traits. *Nat Genet* 1997; **17**: 387–392.
5 Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ *et al*. Finding the missing heritability of complex diseases. *Nature* 2009; **461**: 747–753.
6 Rappaport SM, Smith MT. Environment and Disease Risks. *Science* 2010; **330**: 460–461.
7 Oliver BR, Plomin R. Twins' Early Development Study (TEDS): a multivariate, longitudinal genetic investigation of language, cognition and behavior problems from childhood through adolescence. *Twin Res Hum Genet* 2007; **10**: 96–105.
8 Haworth CMA, Harlaar N, Kovas Y, Davis OSP, Oliver BR, Hayiou-Thomas ME *et al*. Internet cognitive testing of large samples needed in genetic research. *Twin Res Hum Genet* 2007; **10**: 554–563.
9 Boker S, Neale M, Maes H, Wilde M, Spiegel M, Brick T *et al*. OpenMx: an open source extended structural equation modeling framework. *Psychometrika* 2011; **76**: 306–317.
10 Wong B. Design of data figures. *Nat Meth* 2010; **7**: 665–665.
11 Purcell S. Variance components models for gene-environment interaction in twin analysis. *Twin Res* 2002; **5**: 554–571.
12 Hanscombe KB, Trzaskowski M, Haworth CMA, Davis OSP, Dale PS, Plomin R. Socioeconomic Status (SES) and Children's Intelligence (IQ): in a UK-representative sample SES moderates the environmental, not genetic, effect on IQ. *PLoS ONE* 2012; **7**: e30320.
13 Pickett KE, Wilkinson RG. Child wellbeing and income inequality in rich societies: ecological cross sectional study. *BMJ* 2007; **335**: 1080–1080.
14 Wilkinson R, Pickett K. *The spirit level*. Penguin Books: London, 2009.
15 Zink CF, Tong Y, Chen Q, Bassett DS, Stein JL, Meyer-Lindenberg A. Know your place: neural processing of social hierarchy in humans. *Neuron* 2008; **58**: 273–283.
16 Izuma K, Saito DN, Sadato N. Processing of social and monetary rewards in the human striatum. *Neuron* 2008; **58**: 284–294.
17 Plomin R, Asbury K, Dunn J. Why are children in the same family so different? Nonshared environment a decade later. *Can J Psychiatry* 2001; **46**: 225–233.
18 Lederbogen F, Kirsch P, Haddad L, Streit F, Tost H, Schuch P *et al*. City living and urban upbringing affect neural social stress processing in humans. *Nature* 2011; **474**: 498–501.
19 Fonken LK, Xu X, Weil ZM, Chen G, Sun Q, Rajagopalan S *et al*. Air pollution impairs cognition, provokes depressive-like behaviors and alters hippocampal cytokine expression and morphology. *Mol Psychiatr* 2011; **16**: 987–995.
20 Feil R, Fraga MF. Epigenetics and the environment: emerging patterns and implications. *Nat Rev Genet* 2012; **13**: 97–109.
21 Plomin R, Bergeman CS. The nature of nurture: genetic influence on 'environmental' measures. *Behav Brain Sci* 1991; **14**: 373–386.
22 Gottesman II, Gould TD. The endophenotype concept in psychiatry: etymology and strategic intentions. *Am J Psychiatry* 2003; **160**: 636–645.
23 Kendler KS, Neale MC. Endophenotype: a conceptual analysis. *Mol Psychiatry* 2010; **15**: 789–797.
24 Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR *et al*. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 2010; **42**: 565–569.
25 O'Donoghue SI, Gavin A-C, Gehlenborg N, Goodsell DS, Hériché J-K, Nielsen CB *et al*. Visualizing biological data-now and in the future. *Nat Methods* 2010; **7**(3 Suppl): S2–S4.
26 Fox P, Hendler J. Changing the equation on scientific data visualization. *Science* 2011; **331**: 705–708.
27 Scott FJ, Baron-Cohen S, Bolton P, Brayne C. The CAST (Childhood Asperger Syndrome Test): preliminary development of a UK screen for mainstream primary-school-age children. *Autism* 2002; **6**: 9–31.
28 Williams J, Scott F, Stott C, Allison C, Bolton P, Baron-Cohen S *et al*. The CAST (Childhood Asperger Syndrome Test): test accuracy. *Autism* 2005; **9**: 45–68.
29 Conners CK, Sitarenios G, Parker JDA, Epstein JN. The Revised Conners' Parent Rating Scale (CPRS-R): factor structure, reliability, and criterion validity. *J Abnormal Child Psychol* 1998; **26**: 257–268.
30 Angold A, Costello E, Messer S, Pickles A, Winder F, Silver D. The development of a short questionnaire for use in epidemiological studies of depression in children and adolescents. *Int J Methods Psychiatr Res* 1995; **5**: 1–12.
31 Goodman R. The Strengths and Difficulties Questionnaire: a research note. *J Child Psychol Psychiatry* 1997; **38**: 581–586.
32 Goodman R. Psychometric properties of the strengths and difficulties questionnaire. *J Am Acad Child Adolesc Psychiatry* 2001; **40**: 1337–1345.
33 Frick P, Hare R. *Antisocial Process Screening Device*. Multi Health Systems: Toronto, 2001.
34 Vitacco MJ, Rogers R, Neumann CS. The Antisocial process screening device: an examination of its construct and criterion-related validity. *Assessment* 2003; **10**: 143–150.

Supplementary Information accompanies the paper on the Molecular Psychiatry website (http://www.nature.com/mp)