nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-18418-4 OPEN

Dutch population structure across space, time and GWAS design

Ross P. Byrne[1✉], Wouter van Rheenen[2], Project MinE ALS GWAS Consortium*, Leonard H. van den Berg[2], Jan H. Veldink[2] & Russell L. McLaughlin[1✉]

Previous genetic studies have identified local population structure within the Netherlands; however their resolution is limited by use of unlinked markers and absence of external reference data. Here we apply advanced haplotype sharing methods (ChromoPainter/fineSTRUCTURE) to study fine-grained population genetic structure and demographic change across the Netherlands using genome-wide single nucleotide polymorphism data (1,626 individuals) with associated geography (1,422 individuals). We identify 40 haplotypic clusters exhibiting strong north/south variation and fine-scale differentiation within provinces. Clustering is tied to country-wide ancestry gradients from neighbouring lands and to locally restricted gene flow across major Dutch rivers. North-south structure is temporally stable, with west-east differentiation more transient, potentially influenced by migrations during the middle ages. Despite superexponential population growth, regional demographic estimates reveal population crashes contemporaneous with the Black Death. Within Dutch and international data, GWAS incorporating fine-grained haplotypic covariates are less confounded than standard methods.

[1] Smurfit Institute of Genetics, Trinity College Dublin, Dublin D02 DK07, Republic of Ireland. [2] Department of Neurology and Neurosurgery, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht 3584 CX, The Netherlands. *A full list of members and their affiliations appears in the Supplementary Information. ✉email: rbyrne5@tcd.ie; mclaugr@tcd.ie

The Netherlands is a densely populated country on the northwestern edge of the European continent, bounded by Germany, Belgium and the North Sea. The country is divided into twelve provinces and has a complex demographic history, with occupation by several Germanic peoples since the collapse of the Roman Empire, including the Frisians, the Low Saxons and the Franks. Over 17 million individuals now inhabit this relatively small region (41,500 km^2), making it one of the most densely populated countries in Europe. Despite its small geographical size, previous genetic studies of the people of the Netherlands have demonstrated coarse population structure that correlates with its geography, as well as apparent heterogeneity in effective population sizes across provinces[1,2]. These observations suggest that the demographic past of the Dutch population has left residual signatures in its present regional genetic structure; however, this has not been fully explained in the context of neighbouring populations and thus far the use of unlinked genetic markers have limited the resolution at which this structure can be described. This resolution limit also confines the extent to which the confounding effects of population structure can be controlled in genomic studies of health and disease such as genome-wide association studies (GWAS). As these studies continue to seek ever-rarer genetic variation with ever-increasing cohort sizes, intricate understanding and fine control of population structure is becoming increasingly relevant, but increasingly challenging[3].

Recent studies have showcased the power of leveraging shared haplotypes to uncover and characterise previously unrecognised fine-grained genetic structure within populations, yielding novel insights into the demographic composition and history of Britain and Ireland[4–7], Finland[8], Japan[9], Italy[10], France[11] and Spain[12]. Haplotype sharing has also revealed genetic affinities between populations[13], enabling inference of historical admixture events using modern populations as proxies for ancestral admixing sources[14]. Furthermore, geographic information can be integrated to model genetic similarity as a function of spatial distance[15] to infer demographic mobility within or between populations; one approach uses the Wishart distribution to estimate and map a surface of effective migration rates based on deviations from a pure isolation by distance model[16], allowing migrational cold spots to be inferred which may derive from geographical boundaries such as rivers and mountains. Almost half of the area of the Netherlands is reclaimed from the sea and its contemporary land surface is densely subdivided by human-made waterways and naturally-occurring rivers, including the Rhine (Dutch: *Rijn*), Meuse (*Maas*), Waal and IJssel. These rivers have been speculatively linked to genetic differentiation between northern and southern Dutch subpopulations in previous work[1]; however the explicit relationship between Dutch genetic diversity and movement of people within the Netherlands has not been directly modelled.

The Dutch have previously received special interest as a model population[1,2] and form a major component of substantial ongoing efforts to better understand human health, disease, demography and evolution. For example, at the time of writing, over 10% of all studies listed in the NHGRI-EBI genome-wide association study (GWAS) catalogue[17] include the Netherlands in their "Country of recruitment" metadata. As well as offering insights into demography and human history, refined population genetic studies are important to identify and adequately control confounding effects in genomic studies of health and disease, especially if spatially structured environmental factors contribute substantially to variance in phenotype, which in particular impacts rare variants[18].

In this study, we harness shared haplotypes to examine the fine-grained genetic structure and demography of the Netherlands. We show that Dutch population structure is more granular than previously recognised, and is ancient and persistent over time. The strength and stability of the observed structure appears to be tied to the relationship of the Netherlands to neighbouring lands and to its own internal geography, and has likely been shaped over history by migration, but preserved in recent generations by enduring sedentism of genetically similar individuals within regions. We observe genetic evidence of regional population crashes during the Black Death and a countrywide population surge in the 17th century. Finally, we show that the complex genetic structure observed demonstrably confounds GWAS; however, through analysis of the Netherlands and more extensive international data[19], we demonstrate that using shared haplotypes as GWAS covariates significantly reduces this confounding over standard single-marker methods.

Results

The genetic structure of the Dutch population. We mapped the haplotypic coancestry profiles of 1626 Dutch individuals using ChromoPainter[20] and clustered the resulting matrix using fineSTRUCTURE[20], identifying 40 genetic clusters at the highest level of the hierarchical tree which segregated with geographical provenance. We explored the clustering from the finest ($k = 40$) to the coarsest level ($k = 2$), settling on $k = 16$ as it captured the major regional splits sufficiently with little redundancy. Clusters at this level were robustly defined by total variation distance (TVD; $p < 0.001$) and fixation index (F_{ST}; Fig. 1a); remarkably, some F_{ST} values between particularly differentiated Dutch clusters were comparable in magnitude to estimates between European countries (calculated using data from ref. [21]; Supplementary Table 1). Some clusters had expansive geographical ranges (for example NHFG, representing individuals from North Holland, Friesland and Groningen), while others neatly distinguished populations on a sub-provincial level (for example, NBE and NBW, representing east and west regions of North Brabant). For visualisation we projected the ChromoPainter coancestry matrix in lower dimensional space using principal component analysis (PCA; Fig. 1b) and assigned cluster labels based on majority sampling location (available for 1422 individuals), arranging neighbouring and genetically similar clusters into cluster groups, as with previous work[6]. The first principal component (PC) of coancestry followed a strong north-south trend (latitude vs mean PC1 per town $r^2 = 0.52$; $p = 6.8 \times 10^{-72}$) with PC2 generally explained by a west-east gradient (longitude vs mean PC2 per town $r^2 = 0.29$; $p = 3.4 \times 10^{-33}$). Further PCs demonstrated more complex relationships with geography (Supplementary Fig. 1).

As previously observed in different populations[6], the distribution of individuals in this genetic projection generally resembled their geographic distribution (Fig. 1c), with some exceptions. For example, North Brabant is geographically further north than Limburg, but is further separated by PC1 from northern clusters. We explored the possibility that this could instead be explained by relative ancestral affinities to neighbouring lands by modelling the genome of each Dutch individual as a linear mixture of European sources (obtained from ref. [21]) using ChromoPainter, retaining source groups that best matched Dutch individuals for at least 5% of the genome[4] (Fig. 2). The resulting profiles of German, Belgian and Danish ancestries were significantly autocorrelated (p_{DE}, $p_{BE} < 0.0001$; $p_{DK} < 0.001$; Moran's I and Mantel's test) and spatially arranged along geographical directions S66°W, N73°E and S73°E respectively, approximately corresponding to declining ancestry gradients directed away from the German and Belgian borders and the North Sea boundary (Fig. 2; $r^2_{DE} = 0.31$; $r^2_{BE} = 0.35$; $r^2_{DK} = 0.12$; $p_{DE} = 9.4 \times 10^{-119}$; $p_{BE} = 2.7 \times 10^{-133}$; $p_{DK} = 1.1 \times 10^{-39}$). The spatial distribution of French ancestry was comparatively

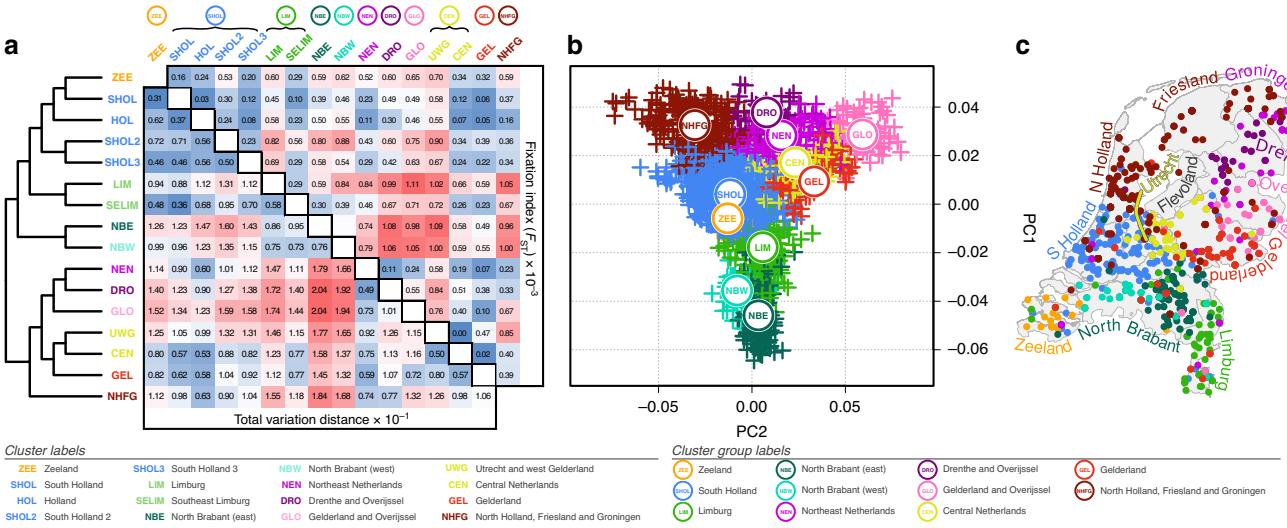

Fig. 1 The genetic structure of the people of the Netherlands. a fineSTRUCTURE dendrogram of ChromoPainter coancestry matrix showing clustering of 1626 Dutch individuals based on haplotypic similarity. Associated total variation distance (TVD) and fixation index statistics between clusters are shown in the matrix. Permutation testing of TVD yields $p < 0.001$ for all cluster pairs, indicating that clustering is non-random. Cluster labels derive from Dutch provinces and are arranged into cluster groups for genetically and geographically similar clusters (circled labels). **b** The first two principal components (PCs) of ChromoPainter coancestry matrix for all individuals analysed. Points represent individuals and are coloured and labelled by cluster group. **c** Geographical distribution of 1422 sampled individuals, coloured by cluster groups defined in **a**. Labels represent provinces of the Netherlands. Map boundary data from the Database of Global Administrative Areas (GADM; https://gadm.org).

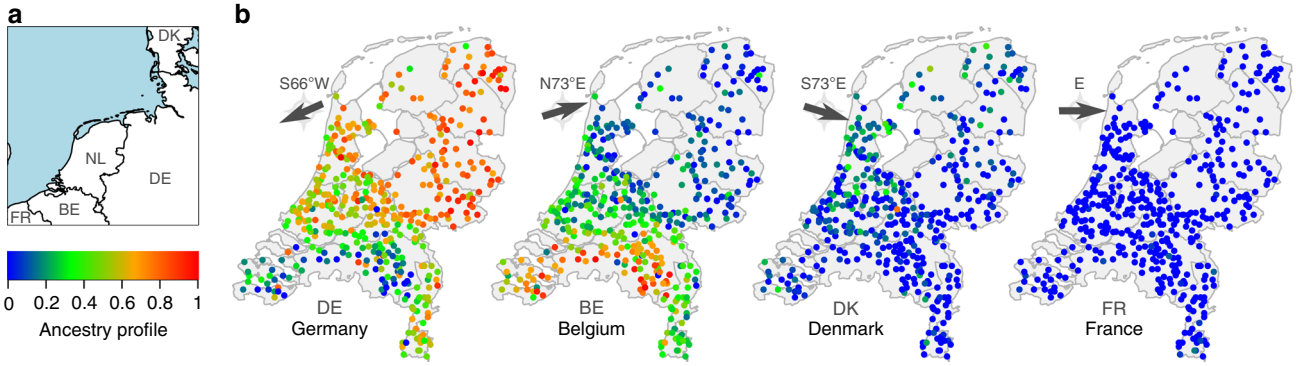

Fig. 2 The ancestry profile of the Netherlands. a The Netherlands and its geographical relationship to neighbouring lands. **b** German, Belgian, Danish and French haplotypic ancestry profiles for 1422 Dutch individuals. Arrows indicate the predominant directions along which the ancestry gradients are arranged across the Netherlands. Map boundary data from the Database of Global Administrative Areas (GADM; https://gadm.org) and Natural Earth (https://naturalearthdata.com).

uniform, with only a modest correlation due east ($r^2_{FR} = 0.014$; $p_{FR} = 9.5 \times 10^{-6}$). The general trend across the Netherlands was thus of complementary Belgian and German ancestral affinities, decaying with distance from the respective borders. North Brabant, however, showed a greater Belgian profile than Limburg, despite similar, substantial Belgian frontiers in both Dutch provinces. Conversely, the German ancestry profile of Limburg greatly exceeded that of North Brabant, reflecting its 200-kilometre border with Germany and centuries of consequent demographic contact and likely genetic admixture.

Genome flux and stasis in the Netherlands. To explore temporal trends in Dutch population structure we called genomic segments of pairwise identity-by-descent (IBD) using RefinedIBD[22]. An IBD haplotype sharing matrix is conceptually similar to a ChromoPainter coancestry matrix[23], but trades some sensitivity to be more explicitly interpretable. As IBD segment length is inversely related to age[24,25], different length intervals can inform on structure at different time depths. Total pairwise IBD between Dutch individuals mirrored the structure observed with ChromoPainter (Fig. 3a), with 8 distinct clusters identified in the IBD sharing matrix that broadly segregated with geography and recapitulated some of the important splits obtained from fineSTRUCTURE, most strikingly the west-east split in North Brabant. Decomposing total IBD by centiMorgan (cM) length into short (1–3 cM), medium (3–5 cM) and long (5–7 cM) bins, we observed stability over time of north-south structure and the emergence of west-east structure embedded in 3–5 cM segments (Fig. 3b), corresponding to an expected time depth around 1120 years ago[25]. As this date and the structure observed is dependent on the (arbitrary) thresholds set for IBD segment length bins, we have also provided an interactive environment in which Dutch population structure can be explored across a range of IBD segment bins (http://bioinf.gen.tcd.ie/ctg/nlibd).

Although these observations could potentially be biased by power to detect population structure in longer and shorter bins, the temporally volatile west-east structure contrasts with the

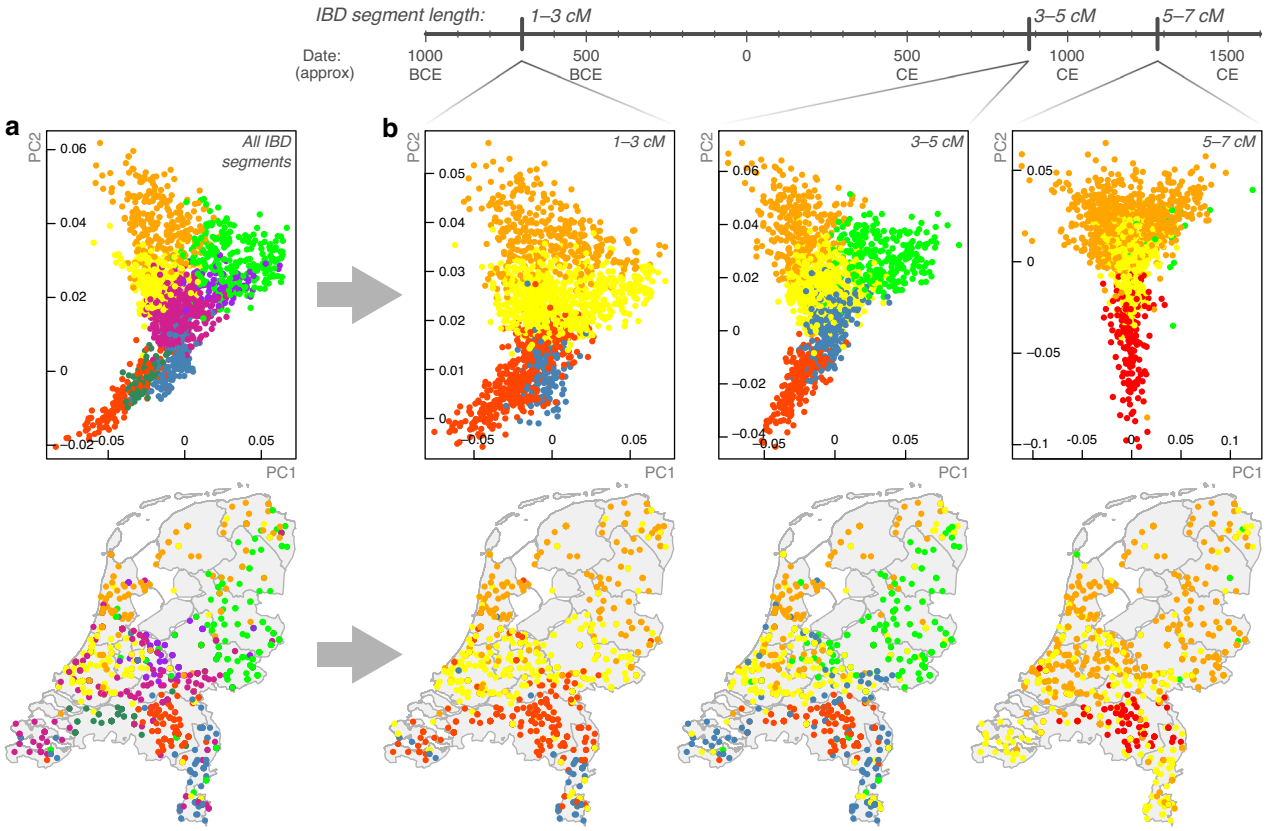

Fig. 3 The changing genomic structure of the Dutch population over time. a Principal component (PC) analysis of pairwise total identity-by-descent (IBD) for 1626 Dutch individuals (top) and their geographical provenance (bottom). Points represent individuals and are coloured by cluster assignment (mclust on pairwise IBD matrix). **b** PCs (top) and geographical provenance (bottom) for pairwise sharing of 1–3, 3–5 and 5–7 centiMorgan (cM) IBD segments, corresponding to point estimates of expected time depths at ~2700, 1120 and 720 years ago, respectively. Time depths for IBD segment bins have wide distributions[25]; expected values presented here should be interpreted as a guide only and the changing west-east structure over time does not necessarily reflect (for instance) a precisely-timed admixture event. Map boundary data from the Database of Global Administrative Areas (GADM; https://gadm.org).

Table 1 GLOBETROTTER date and source estimates for admixture into the Netherlands.

Cluster group	Conclusion	Minor	Major	Prop	Date CE	95% CI CE	*p*
SHOL	One-date multiway	SPA-FRA(2)	GER(5)	0.25	1169	1086–1244	0
ZEE	One-date-multiway	FRA(8)	GER(5)	0.4	1172	771–1773	0
NBE	One-date-multiway	FRA(8)	GER(5)	0.4	1085	939–1262	0
NBW	One-date-multiway	GER(5)	BEL(5)	0.34	1013	668–1383	0
NEN	One-date	SPA-FRA(2)	GER(5)	0.19	1172	925–1364	0
DRO	One-date-multiway	FRA(8)	GER(5)	0.16	1390	1116–1932	0
GLO	One-date	SPA-FRA(2)	GER(5)	0.14	1128	893–1306	0
CEN	One-date	SPA-FRA(2)	GER(5)	0.18	1049	854–1244	0
GEL	One-date	SPA-FRA(2)	GER(5)	0.17	1189	1046–1391	0
NHFG	One-date	GER(9)	DEN(5)	0.36	1060	759–1290	0
LIM	One-date	ITA(8)	GER(5)	0.34	1162	1044–1351	0
ALL	One-date	SPA-FRA(2)	GER(5)	0.25	1088	1004–1111	0

Minor and **Major** represent inferred proxy admixing sources. **Prop** represents estimated minor admixture proportion. Admixing sources are derived from ChromoPainter/fineSTRUCTURE clustering of 4514 European reference individuals (Methods); labels represent principal country of origin (SPAin, FRAnce, GERmany, BELgium, DENmark) with cluster numbers arbitrarily assigned within countries. Example coancestry curves are shown in Supplementary Fig. 2.

stability and persistence of old north-south structure and possibly represents a genomic signature of historical demographic flux in the region and its surrounding lands. With this in mind, we investigated possible admixture from outside demographic groups using GLOBETROTTER[14] with 4514 European individuals[21] representing modern proxies for admixing sources. Across the Dutch sample, significant admixture dating to 1088 CE (95% CI 1004–1111 CE) was inferred with the major contributing source best modelled by modern Germans and the minor source best modelled by southern European groups (France, Spain) (Table 1). This is supported by single-marker ADMIXTURE component estimates showing that the Netherlands has the closest profile to Germanic groups (Supplementary Fig. 3) and is consistent with the ancestry profile gradients detailed in Fig. 2. The timing of the inferred 11th century event was stable across Dutch fineSTRUCTURE clusters (to varying

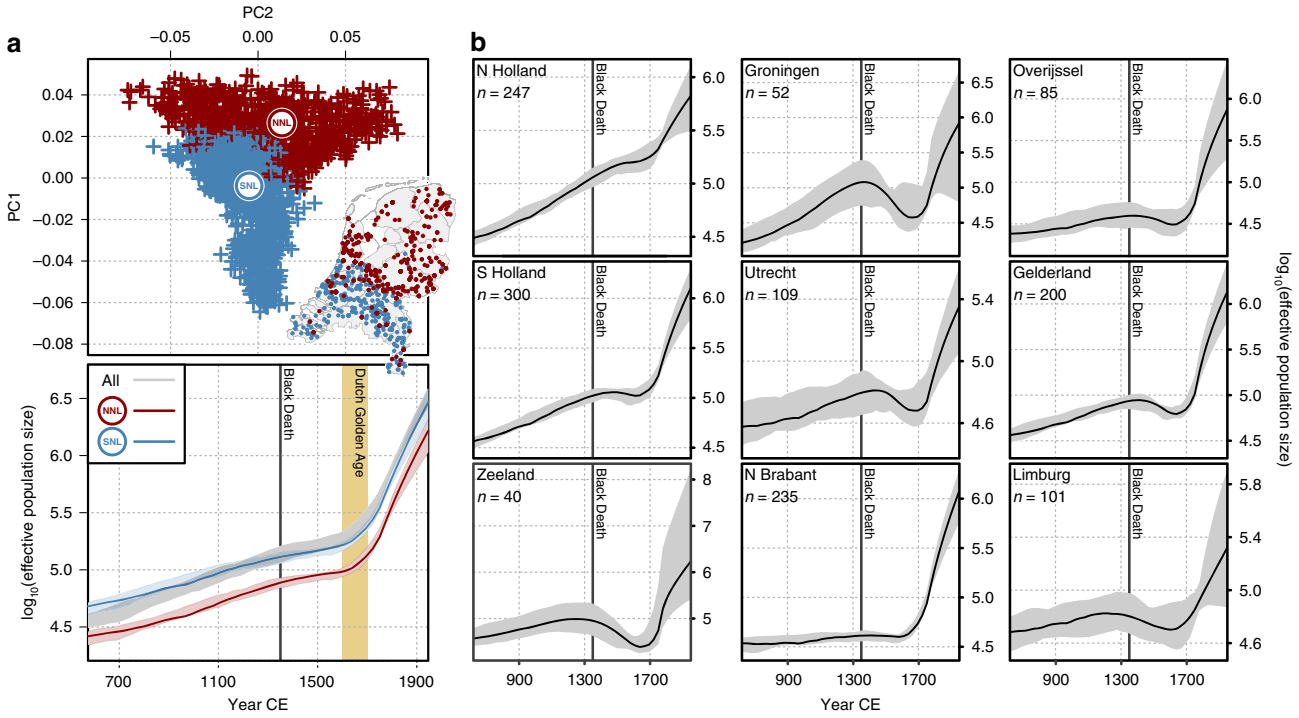

Fig. 4 Dutch effective population size over time. a Historical change in effective population size (N_e) over the past 50 generations for all Dutch individuals and subsets of northerners and southerners. The top plot shows the principal components of ChromoPainter coancestry coloured by the first ($k = 2$) fineSTRUCTURE split, which separates the Dutch population into northern (NNL) and southern (SNL) genetic clusters; inset shows geographical distribution of these individuals. The bottom plot shows growth in effective population size countrywide or per fineSTRUCTURE cluster over the past 50 generations. **b** Historical N_e trajectories for individual Dutch provinces with more than 40 individuals sampled. For both (**a**, **b**), curves show point estimates for N_e bounded by a 95% CI estimated from 80 bootstraps of the data (note this is not symmetrically distributed around the point estimates) and assume 28 years per generation and mean year of birth at 1946 CE. Map boundary data from the Database of Global Administrative Areas (GADM; https://gadm.org).

degrees of confidence), suggesting that the signal represents an important period in the establishment of the modern Dutch genome (Table 1); however, given the state of demographic flux in Europe at the time, its exact historical correlate is open to interpretation. Notably, a significant admixture event with a major Danish source was inferred between 759 and 1290 CE in the NHFG cluster group (representing Dutch northern seaboard provinces); this period spans a historical period of recorded Danish Viking contact and rule in northern Dutch territories.

In addition to influence from outside populations, the population structure detailed in Figs. 1 and 3 has likely been shaped by independent demographic histories within the Netherlands. In support of this, we noted that short (1–2 cM) IBD segments shared between northern clusters and provinces outnumbered those shared between southern clusters and provinces (Supplementary Fig. 4), and, as observed previously[2], northern provinces shared more short segments with southern provinces than southern provinces shared amongst themselves. Together, these results suggest that the north had a smaller ancestral effective population size (N_e) than the south and is probably derived from an ancient or historical founder event forming the northern population from a subset of southerners. We formally characterised ancestral trajectories in N_e between the north and the south of the Netherlands using the nonparametric method IBDNe[26] for the entire Dutch sample and two subsamples representing the principal fineSTRUCTURE north/south split (Fig. 4a), retaining a random sample of 641 individuals from each group. We also characterised historical N_e within individual Dutch provinces for which genotypes for more than 40 individuals were available. Countrywide, N_e has grown super-exponentially over the past 50 generations in the Netherlands (Fig. 4a) and has been consistently lower in the north than the south. Despite this, the pattern of growth in northern and southern groups was identical, with a steady exponential growth up to around 1650 CE, when a major uptick in growth rate was observed. This corresponds to a period of substantial economic development in the Netherlands over the 17th century known to historians as the Dutch Golden Age. Preceding this period, historical N_e estimates for the entire country and for northern/southern groups showed only a modest response to the Black Death (*Yersinia pestis* plague pandemic) of the 14th century which claimed up to 60% of Europe's population[27]. Conversely, N_e estimation within individual Dutch provinces revealed a much more detectable impact of the Black Death (Fig. 4b).

Genomic signatures of Dutch mobility. We noted that long (>7 cM) IBD segments, which capture recent shared ancestry, were almost always shared within genetic clusters (and provinces), and rarely between (Supplementary Fig. 4). This indicates a propensity for genetically similar individuals (relatives) to remain mutually geographically proximal, suggesting a degree of sedentism that has likely influenced Dutch population structure over time. It has also previously been argued that genetic structure in the Netherlands may be partially rooted in geographic obstacles imposed by the country's major waterways[1] so we explicitly modelled genetic similarity as a function of geographic distance using EEMS[16] to infer migrational hot and cold spots (Fig. 5). The resulting effective migration surface showed several apparent barriers to gene flow, the strongest and most contiguous of which runs in an east-west direction across the Netherlands overlapping the courses of the Rhine, Meuse and Waal rivers. This inferred migrational boundary also approximately

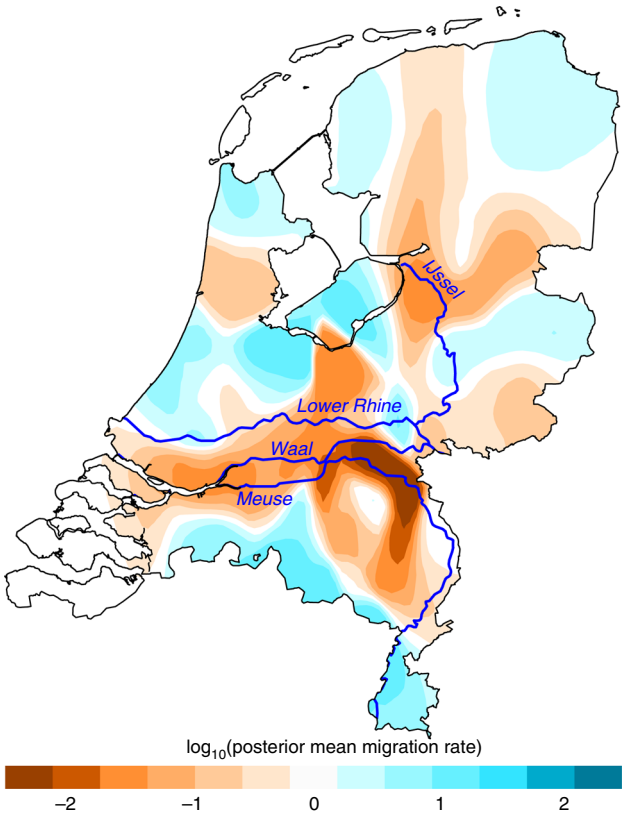

Fig. 5 The effective migration surface of the Netherlands. Contour map shows the mean of 10 independent EEMS posterior migration rate estimates between 800 demes modelled over the land surface of the Netherlands. A value of 1 (blue) indicates a tenfold greater migration rate over the average; −1 (orange) indicates tenfold lower migration than average. The courses of major rivers are included to highlight their correlation with migrational cold spots. Map boundary data from the Database of Global Administrative Areas (GADM; https://gadm.org); river course data from Natural Earth (https://www.naturalearthdata.com).

corresponds to the geographical division determining the principal fineSTRUCTURE split between northern and southern Dutch populations (Fig. 4a) as well as the geographical boundaries between clusters inferred from ancient IBD segments (Fig. 3b), suggesting that these rivers have been a historically persistent determinant of Dutch population structure.

GWAS confounding by fine-grained structure. As population structure confounds GWAS (for example due to stratification of cases and controls between subpopulations), we investigated the extent to which haplotype sharing captures confounding structure in a Dutch sample of 1963 cases of amyotrophic lateral sclerosis (ALS) and 2774 controls from a recent multi-population GWAS for ALS[19]. PCs of the haplotypic ChromoPainter coancestry matrix for these 4737 individuals explained substantially more variance in ALS phenotype than PCs calculated from single nucleotide polymorphism (SNP) genotypes alone, indicating latent structure captured by ChromoPainter that is stratified between cases and controls (Fig. 6a). To estimate the extent to which this stratified structure confounds GWAS we calculated case-control association statistics using a logistic model covarying for either 20 ChromoPainter PCs or 20 SNP PCs and estimated the linkage disequilibrium (LD) score regression intercepts for both sets of resulting summary statistics. An intercept higher than 1 indicates confounding in the GWAS; Fig. 6a shows that GWAS statistics calculated with ChromoPainter PCs as covariates are less confounded than statistics using SNP PCs, albeit with overlapping confidence intervals for the relatively small Dutch sample. To more adequately represent the large-scale multi-population data typically used in modern GWAS, we extended our analysis to the full ALS case-control dataset from which the Dutch data derive[19], including 35,755 individuals from twelve European countries and the USA. For computational tractability, instead of ChromoPainter we used PBWT-paint (https://github.com/richarddurbin/pbwt), a scalable approximate haplotype painting method based on the positional Burrows-Wheeler transform[28]. When run on our original Dutch dataset of 1626 individuals, the structure rendered by PBWT-paint was almost identical to ChromoPainter ($r^2_{PC1} = 0.99$; $r^2_{PC2} = 0.98$; Supplementary Fig. 5), indicating its suitability for this analysis. PBWT-paint captured pervasive global and local structure in the multi-population GWAS data that both separated and subdivided countries (Fig. 6b). Top PCs of PBWT-paint coancestry explained substantially more variance in phenotype than SNP PCs and GWAS statistics including PBWT-paint PCs as covariates were significantly less confounded than statistics corrected by SNP PCA (Fig. 6a, LD score regression intercepts).

Discussion

The genomes of modern humans contain a detailed record of the intricate histories that shaped them. Genomic signatures of these histories are often reflected in present-day population structure and have the potential to confound genomic studies of health and disease through stratification across phenotypic categories. Here, we have studied the Netherlands as a model population, harnessing information from shared haplotypes and recent developments in spatial modelling to gain intricate insights into the geospatial distribution and likely origin of Dutch population genetic structure. The structure identified through shared haplotypes is surprisingly strong; some Dutch genetic clusters identified this way are more mutually distinct (by F_{ST}) than whole European countries. We have also introduced a novel use of length-binned IBD sharing combined with PCA and Gaussian mixture model-based clustering to characterise changing population structure over time, revealing transient genetic structure layered over strong and stable north-south differentiation in the Netherlands. This is contextualised by somewhat distinct demographic histories between genetic groups in the Netherlands, with consistently lower N_e in the north than the south. A potential source of the north-south differentiation is impaired migration across the east-west courses of the Rhine, Meuse and Waal, which effectively separate southern Dutch populations from the north. The population structure observed in the Netherlands is especially remarkable when considered in terms of the country's size and extensive infrastructure; notably Denmark, which is roughly equal in geographical area, is genetically homogeneous, forming only a single cluster when interrogated using fineSTRUCTURE[29], despite its island-rich geography. Both the United Kingdom and Ireland also exhibit at least one large indivisible cluster constituting a large fraction of the population[4–6], however no extraordinarily large clusters dominate the Dutch sample. Mean F_{ST} between Dutch clusters also greatly outmeasures that observed between Irish clusters, suggesting that the extent of population differentiation is higher in the Netherlands, despite Dutch land area being less than half that of the island of Ireland.

While coarse geographical trends in Dutch genetic structure have previously been described using single-marker PCA[1], our use of shared haplotypes reveals structure at a much higher resolution, differentiating subpopulations between, and sometimes within, provinces (Fig. 1). As a striking example, individuals from the east and west of North Brabant (NBE and NBW in

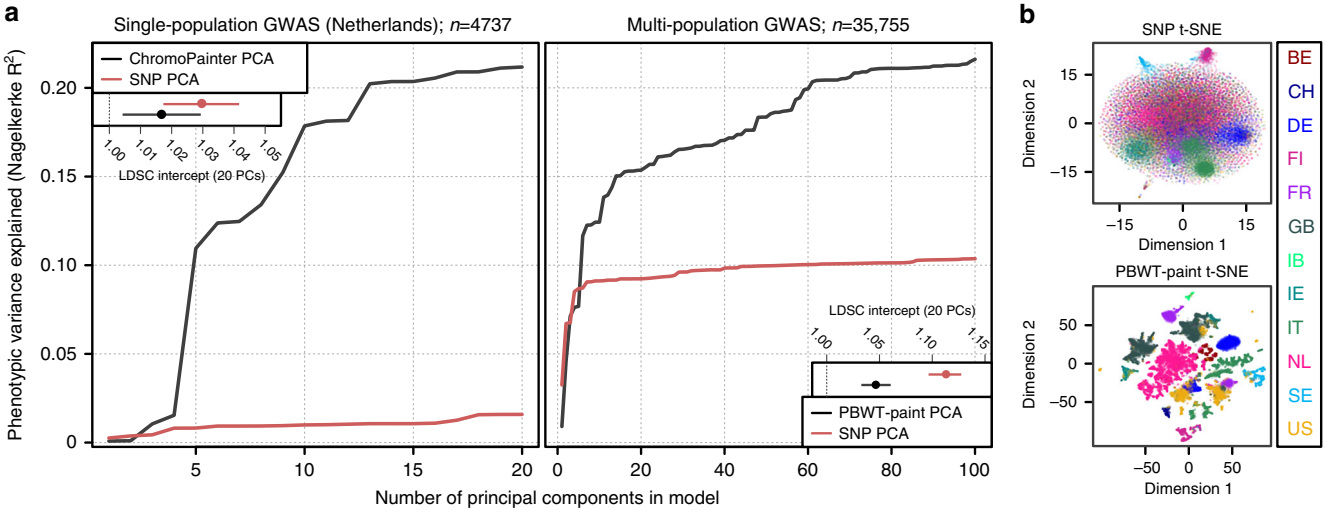

Fig. 6 Fine-grained population structure and genome-wide association study (GWAS) confounding. a Variance in phenotype (amyotrophic lateral sclerosis) explained by principal components (PCs) for a single-population Dutch GWAS (left) and a multi-population GWAS (right). Insets show linkage disequilibrium score regression (LDSC) intercept terms (a summary estimate of GWAS confounding) when the first 20 single nucleotide polymorphism (SNP)-based PCs (SNP PCA) or the first 20 haplotype-based PCs (ChromoPainter/PBWT-paint PCA) are included as GWAS covariates. **b** Summary visualisations (t-distributed stochastic neighbour embedding, t-SNE) of local and global structure in the multi-population GWAS based on SNP genotypes (top) or haplotype sharing inferred using the scalable PBWT-paint chromosome painting algorithm (bottom). Individuals are coloured by country of origin; labels (right) follow ISO 3166-1 country codes, except IB, which was labelled Iberia (containing Spanish and Portuguese data) in the original GWAS dataset. PCA, principal component analysis; PBWT, positional Burrows-Wheeler transform.

Fig. 1) are mutually genetically distinguishable and are more distinct from clusters to their north than Limburg, despite being geographically closer. This deviation from haplotype sharing mirroring geography appears to be driven by strong genetic affinity to Belgium (Fig. 2), reflecting a long history of demographic and sovereign overlap across a 100 km frontier spanning the modern Dutch-Belgian border. In contrast, the majority of ancestral influence in Limburg, which also shares a substantial border with Belgium, is equally split between Belgium to the west and Germany to the east. Notably, the Belgian border with the south of Dutch Limburg is almost entirely described by the course of the Meuse, which may have acted as a historical impediment to migration, thus distinguishing individuals in this region genetically. This is reflected in IBD clustering, in particular the distinction of southern Limburgish individuals from the rest of the Netherlands in short (1–3 cM) segments, which otherwise only describe coarse north-central-south structure (Fig. 3). Future work explicitly modelling Dutch-Belgian and Dutch-German frontiers using additional Belgian and German genetic data with associated geography will resolve the historical and present-day role of the Meuse in distinguishing distinct population clusters in the south of the Netherlands.

Similarly to North Brabant, groups of individuals in North and South Holland show significant genetic separation despite mutual geographic proximity. While we have chosen to group the four South Holland clusters for visual brevity in Fig. 1, they are robustly distinct by TVD permutation analysis ($p < 0.001$), indicating that significant population differentiation exists even within South Holland. Migration and admixture in the highly urbanised *Randstad* has been proposed as a driver of genetic diversity and loss of geographic structure in this region[1]; the overlaid geographical distribution of regional ancestry profiles (Fig. 2) for this area lends support to this hypothesis. However, the geographical ranges of the four South Holland clusters are somewhat independent (Supplementary Fig. 6), indicating that some degree of genetic structure has survived this urbanisation. Previous studies have highlighted the correlation between decreasing autozygosity and increased urbanisation[30]; future work leveraging the ChromoPainter/fineSTRUCTURE framework coupled with length-binned IBD and Gaussian mixture model-based clustering will more explicitly delineate the interplay between urbanisation and population structure over time. To this end, highly urbanised areas such as the *Randstad* will be particularly informative.

The principal fineSTRUCTURE split in the Netherlands describes north-south genetic differentiation (Fig. 1) that is strong and persistent over time (Fig. 3). We hypothesised that this reflects partially independent demographic histories so we estimated ancestral N_e for northern (NNL) and southern (SNL) Dutch fineSTRUCTURE populations, revealing superexponential growth in both populations with a sudden increase in rate during the 17th century (Fig. 4a). Historical N_e follows the same approximate trajectory for both populations but is consistently lower for the northern cluster, corroborating previous observations of increased homozygosity in northern Dutch populations[1] and consistent with a model of northerners representing a founder isolate from southerners (although a more complex demographic model may better explain these observations)[1,2]. The apparent absence of N_e decline in 14th-century Netherlands initially hints at the possibility that the Black Death had a weaker impact in the region than elsewhere in Europe; although this agrees with the views of some historians, it is hotly debated by others[31]. Per province, however, most N_e estimates display a prominent dip at this time (Fig. 4b), suggesting that merging non-randomly mating subpopulations into a countrywide group (Fig. 4a) artificially inflates diversity, thus smoothing over any population crash following the Black Death. Population structure is thus important when estimating N_e and trends countrywide and in NNL and SNL clusters (Fig. 4a) should be interpreted carefully: it is possible that a substantial population crash brought about by the Black Death might have had only a marginal impact on the overall effective size of the breeding population in these merged groups. Indeed, the rate of exponential growth in countrywide N_e (Fig. 4a) is marginally shallower in the 10 generations

following the Black Death (0.024; 95% CI 0.0235–0.0251) compared to the 10 generations prior (0.017; 95% CI 0.016—0.018), indicating enduring strain on the overall Dutch population prior to its recovery in the 17th century.

Previous works have hinted that north-south genetic differentiation in the Netherlands may have been facilitated by cultural division between the predominantly Catholic south and the Protestant north[1]. Given that the north-south structure observed in 1–3 cM IBD bins (expected time depth ~700 BCE) greatly precedes different forms of Christianity (Fig. 3), our data support a model in which the Protestant Reformation of the 16th and 17th centuries exploited pre-existing demographic subdivisions, leading to correlation between distinct cultural affinities and clusters of genetic similarity which has potentially been further strengthened by assortative mating among religious groups[32]. Geographical modelling supports the role of migrational boundaries in establishing and maintaining this population substructure, especially rivers (Fig. 5). A substantial belt of low inferred migration runs across the Netherlands, corresponding closely to the roughly parallel east-west courses of the Lower Rhine, Waal and Meuse rivers and correlating with the geographical boundary of the principal north-south fineSTRUCTURE split. Absolute assignment of causality to these geographical correlates is, however, not possible and, given the dense network of waterways in the Netherlands, could be misleading. For example, a strong migrational cold spot in the east of the Netherlands runs parallel to the IJssel (Fig. 5), but could potentially be better explained by the course of the Apeldoorn Canal, a politically fraught waterway constructed in the early 19th Century. Similarly, a cold spot in the northwest directly overlays the North Sea Canal (completed in 1876). As both of these are human-made waterways, it is not certain whether their courses are consequences or determinants of low movement of people across their paths.

As well as internal geography, outside populations have also played an important and significant role in the establishment of population structure in the Netherlands (Fig. 2; Table 1); however the variety and extent of demographic upheaval and mobility of European populations over history obscure the likely historical provenance of most inferred admixture signals. As an important exception, however, ancestry profiles show a small but significant contribution of Danish haplotypes in the north and west of the Netherlands, a possible vestige of Viking raids in coastal areas in the 9th and 10th centuries. This is corroborated by an inferred GLOBETROTTER single-date admixture event in the NHFG (North Holland, Friesland and Groningen) cluster (Fig. 1) between 759 and 1290 CE with Danish haplotypes as a major admixing source (Table 1). The demographic legacy of more than a century of Danish Viking raids and settlement in the Netherlands has been the subject of some debate; from our data, it appears that the modern Dutch genome has indeed been partially shaped by historical Viking admixture. This Danish Viking contact is contemporaneous with a critical period in the establishment of the modern Dutch genome from other outside sources (1004–1111 CE; Table 1), although the precise historical correlates of the admixture events detected in the remaining Dutch regions are less obvious. Future densely sampled ancient DNA datasets from informative time depths in the Netherlands and northwest Europe will enable direct estimation of ancestral population structure, admixture, demographic affinities and effective population sizes, improving precision over the current study which depends on proxy patterns of haplotype sharing between modern individuals. Similarly, regional ancestry and admixture inference are limited by the use of modern proxy populations in place of true ancestral sources; nevertheless, there are ample advantages to the use of modern data, including large sample size and relevance to research on modern human health and disease. In particular, as in our previous work in Ireland[6], samples in the current Dutch dataset were not specifically selected to have pure ancestry in each geographical area (eg all grandparents from the same region[4]) meaning the degree of structure observed is not idealised or exaggerated by sampling, but instead representative of the structure expected in any GWAS that includes Dutch data.

We therefore explored the impact of fine-scale genetic structure described in this study and others[4–12] on GWAS statistics, using the ALS study from which the Dutch data derive as an exemplar trait. Generally, population-based PCs should not predict case/control status (in the absence of any disease-ancestry interaction); if they do, this indicates that (sub)populations are stratified between cases and controls, introducing bias that artificially inflates GWAS statistics. In both Dutch-only and multi-population analyses, fine-scale genetic structure detected by haplotype sharing (ChromoPainter or PBWT-paint) explained substantially more variance in phenotype (ALS case/control status) than standard SNP-only PCA (Fig. 6a). This demonstrates the power of shared haplotypes to simultaneously capture subtle genetic structure within single countries (that is potentially invisible to standard single-marker PCA) along with broader structure between countries and potential cryptic technical artefacts such as platform- or imputation-derived bias. We found that shared haplotypes are effective for controlling GWAS inflation: statistics calculated using haplotype-based PCs as covariates showed lower overall confounding than single marker-based covariates, as measured by LD score regression intercepts (Fig. 6a). In the age of large-scale, single-country and cross-population biobanks, the additional power of haplotype sharing methods to detect fine-scale local population structure will be crucial for ensuring robust GWAS results unconfounded by ancestry. For example, a recent study of latent structure in the UK Biobank demonstrated that a GWAS for birth location returned significant loci even after correction for 40 single-marker PCs[33], suggesting that residual fine-grained population structure may influence other GWAS from this cohort (although others suggest a role for socioeconomically-driven migration in this phenomenon[34]). Ongoing developments in scalable haplotype sharing algorithms such as PBWT-paint will help to address this problem by facilitating the creation of biobank-scale haplotype sharing resources, simultaneously improving studies of human health and disease and enabling large-scale, fine-grained population genetic studies of human demography. Such resources will likely be particularly useful in studies of rare variation, motivating future work exploring the efficacy of such strategies in correcting confounding where rare variation is a factor.

Methods

Data and quality control. We mapped fine-grained genetic structure in the Netherlands using a population-based Dutch ALS case-control dataset ($n = 1626$; subset of stratum sNL3 from a GWAS for amyotrophic lateral sclerosis[19]) and a European reference dataset subsampled from a GWAS for multiple sclerosis[21] (MS; $n = 4514$; EGA accession ID EGAD00000000120 [https://www.ebi.ac.uk/ega/datasets/EGAD00000000120]). 1422 Dutch individuals had associated residential data (hometown at time of sampling) which were used for geographical analyses. For estimating GWAS confounding, we separately analysed the Netherlands on its own using a larger ALS case/control dataset ($n = 4753$; strata sNL1, sNL3 and sNL4 from ref. [19]) and the complete multi-population GWAS dataset[19] ($n = 36{,}052$) from which this Dutch subset was derived. Data handling for estimating confounding is further described under "Estimating GWAS confounding" below. For population structure analyses, we applied quality control (QC) using PLINK v1.9[35]; briefly we removed samples with high missingness (>10%), high heterozygosity (>3 median absolute deviations from median) and single-marker PCA outliers (>5 standard deviations from mean for PCs 1-20). We also filtered out A/T and G/C SNPs and SNPs with minor allele frequency <0.05, high missingness (>2%) or in Hardy Weinberg disequilibrium ($p < 1 \times 10^{-6}$). Before running ChromoPainter/fineSTRUCTURE we retained only one individual from any pair or group that

exhibited greater than 7.5% genomic relatedness ($\hat{\pi}$) and removed SNPs with any missing genotypes as the algorithm does not tolerate missingness or relatedness well. For European reference data we also removed individuals suggested by the QC of the source study[21] and we extracted individuals only of European descent. As this European dataset included MS patients, we filtered out SNPs in a 15 Mb region surrounding the strongly associated HLA locus (GRCh37 position chr6:22,915,594–37,945,593) to avoid bias generated from this association, following previous works. The final Dutch and European reference datasets contained 374,629 SNPs and 363,396 SNPs respectively at zero missingness. The merge of these datasets contained 147,097 SNPs at zero missingness. Data were phased per chromosome with the 1000 Genomes Project phase 3 reference panel[36] using SHAPEIT v2[37] (for ChromoPainter/fineSTRUCTURE) and Beagle v4.1 (for IBD estimation). For these and all subsequent runs of SHAPEIT and ChromoPainter, we used the 1000 Genomes Project Phase 3 genetic map; IBD analyses with Beagle were carried out using the Hapmap phase 2 genetic map[38] as used in the RefinedIBD and IBDNe source papers[22,26]. Both programmes were run with default settings; allele concordance was checked prior to phasing (SHAPEIT: -check; Beagle: conform-gt utility).

fineSTRUCTURE analysis. We used ChromoPainter/fineSTRUCTURE[20] to detect fine-grained population structure using default settings. In brief, each individual was painted using all other individuals (-a 0 0), first estimating N_e and μ (switch rate and mutation rate) with 10 expectation-maximisation (EM) iterations (using all samples and chromosomes), then the model was finally run using these parameter estimates. The fineSTRUCTURE Markov chain Monte Carlo (MCMC) model was then run on the resulting Dutch coancestry matrix with two chains for 3,000,000 burnin and 1,000,000 sampling iterations, sampling every 10,000 iterations. To define European clusters for use in GLOBETROTTER and ancestry profile estimation we instead used 1,000,000 burnin and sampling iterations, sampling every 1000 iterations (due to large sample size). We extracted the state with the maximum posterior probability and performed an additional 10,000 burnin iterations before inferring the final trees using both the climbtree and maximum concordance methods. For all subsequent analyses the maximum concordance tree was used.

Cluster robustness and differentiation. To assess the robustness of clustering in the Dutch data we calculated TVD[4] and F_{ST}. TVD is a distance metric for assessing the distinctness of pairs of clusters, calculated from the ChromoPainter chunklength matrix. TVD is calculated as the sum of the absolute differences between copying vectors for all pairs of clusters, where the copying vector for a given cluster A is a vector of the average lengths of DNA donated to individuals in A by all clusters. Intuitively, the TVD of two clusters reflects distance between those clusters in terms of haplotype sharing amongst all clusters, and is a meaningful method for assessing the effectiveness of fineSTRUCTURE clustering. To assess whether the observed clustering performed better than chance we permuted individuals between cluster pairs (maintaining cluster size) and calculated the number of permutations that exceeded our original TVD score for that pairing of clusters. We used 1000 permutations where possible, and otherwise used the maximum number of unique permutations. P values were calculated from the number of permutations greater than or equal to the observed TVD divided by the total permutations; all p-values were less than 0.001, indicating robust clustering. We generated a TVD tree for clusters from the $k = 16$ fineSTRUCTURE split by merging pairs of clusters with the lowest TVD successively using methods developed in ref. [8], with the goal of providing an alternative representation of cluster relationships that is independent of sample size (Supplementary Fig. 7). The tree was built in k-1 steps, with TVD recalculated at each step from the remaining populations. Branch lengths were scaled proportional to the TVD value of the corresponding pair of populations using adapted code from the original paper[8]. Finally, to assess cluster differentiation independently of the ChromoPainter model, F_{ST} was calculated between Dutch clusters using PLINK 1.9. For this analysis we used the SNP overlap between Dutch and European datasets, pruning for LD (--indep-pairwise 1000 50 0.25) and simultaneously calculating F_{ST} between European countries present in ref. [21] for comparison.

Ancestry profiles. We assessed the ancestral profile of Dutch samples in terms of a European reference made up of 4514 European individuals[21] from Belgium, Denmark, Finland, France, Germany, Italy, Norway, Poland, Spain and Sweden. European samples were first assigned to homogeneous genetic clusters using the fineSTRUCTURE maximum concordance tree[6] to reduce noise in painting profiles. We then modelled each Dutch individual's genome as a linear mixture of the European donor groups using ChromoPainter, and applied ancestry profile estimation method developed in ref. [4] and implemented in GLOBETROTTER[14] (num.mixing.iterations: 0). This method estimates the proportion of DNA which is most closely shared with each individual from each donor group calculated from a normalised ChromoPainter chunklength output matrix, and then implements a multiple linear regression of the form

$$\mathbf{Y}_p = \boldsymbol{\beta}_1\mathbf{X}_1 + \boldsymbol{\beta}_2\mathbf{X}_2 + \ldots + \boldsymbol{\beta}_G\mathbf{X}_G \quad (1)$$

to correct for noise caused by similarities between donor populations. Here, $\mathbf{Y}_p$ is a vector of the proportion of DNA that individual p copies from each donor group, and $\mathbf{X}_g$ is the vector describing the average proportion of DNA that individuals in donor group g copy from other donor groups G, including their own. The coefficients of this equation $\boldsymbol{\beta}_1 \ldots \boldsymbol{\beta}_G$ are thus interpreted as the "cleaned" proportions of the genome that target individual p copies from each donor group, hence the ancestral contribution of each donor group to that individual. The equation is solved using a non-negative-least squares (NNLS) function such that $\boldsymbol{\beta}_g \geq 0$ and the sum of proportions across groups equals 1. We discarded European groups that contributed less than 5% total to any individual, and refit to eliminate noise. We then aggregated sharing proportions across donor groups (genetically homogenous clusters) from the same country to estimate total sharing between an individual and a given country to investigate the regional distribution of sharing profiles. Autocorrelation of ancestry profiles was assessed by Moran's I and Mantel's test (10,000 permutations) in R version 3.2.3. Geographical directions of ancestry gradients were determined by rotating the plane of latitude-longitude between 0° and 360° in 1° steps and finding the axis Y that maximised the coefficient of determination for the linear regression $Y \sim A_c$, where A_c is the aggregated ancestry proportion for country c.

Additionally we compared the ancestry profiles estimated by the NNLS method to those estimated using the recently developed Bayesian algorithm SOURCEFIND[13]. We ran SOURCEFIND on the ChromoPainter output described above using 50,000 burnin and 200,000 MCMC iterations, sampling every 5000 iterations. For each Dutch individual we took the weighted average (weighted by posterior probability) of ancestry estimates with the highest posterior probability taken from 50 independent runs of the algorithm. We aggregated sharing portions across donor groups from the same country to estimate total sharing between an individual and a given country to investigate the regional distribution of sharing profiles. Ancestry gradients generated by each method were regressed against one another to estimate correlation. We report both the results of both NNLS (Fig. 2) and SOURCEFIND (Supplementary Fig. 8) for comparison.

Identity-by-descent analyses. IBD segments were called in phased data using RefinedIBD[22] (default settings) to generate pairwise matrices of total length of IBD shared between individuals for bins of different segment lengths. To identify population structure captured by IBD sharing patterns we performed PCA on these matrices using the prcomp function in R version 3.2.3[39] and clustered the IBD matrices using a Gaussian mixture model implemented in the R package mclust[40]. Plots of model selection are shown in Supplementary Fig. 9. We note that while previous work[23] has shown that IBD matrices underperform the linked ChromoPainter matrix in identifying population structure, they are arguably more interpretable for visualising temporal change as they can be subdivided into cM bins corresponding to different time periods, a feature leveraged by emerging work on local population structure[25]. Patterns in IBD sharing that identify population subgroups in older (shorter) cM bins which are preserved in more recent (longer) bins are interpreted as persistent population structure that has been influenced by mating patterns in old and recent generations. Structure which emerges in a specific cM bin and is lost is likely to reflect transient changes in panmixia that have not necessarily persisted. We approximated the age of segments in a given cM bin using equation s19 from ref. [25], under the assumption that the population is sufficiently large:

$$\lim_{N\to\infty} \mathrm{E}[T|\mu \leq l \leq v] = 75\left(\frac{1}{L_1} + \frac{1}{L_2}\right), \quad (2)$$

where T is the random coalescence time in generations, l is the length of a segment (in base pairs), μ and v are the upper and lower segment length bounds of the interval (in base pairs) and L_2 and L_1 are the upper and lower bounds of the interval rescaled to centiMorgan (i.e. multiplied by $100r$, where r is the recombination rate). For the age estimates given in Fig. 3, we multiplied the expected coalescence time in generations by the approximate human generation time (28 years).

Inferring admixture events. To infer and date admixture events from European sources we ran GLOBETROTTER[14] with the Netherlands dataset as a whole and in individual cluster groups defined from the Dutch fineSTRUCTURE maximum concordance tree (Fig. 1). To define European donor groups we used the European fineSTRUCTURE maximum concordance tree to ensure genetically homogenous donor populations. We used ChromoPainter v2 to paint Dutch and European individuals using European clusters as donor groups (estimating N_e and μ using the weighted average of 10 EM iterations on chromosomes 1, 8, 15 and 20, using all samples). This generated a copying matrix (chunklengths file) and 10 painting samples for each Dutch individual. GLOBETROTTER was run for 5 mixing iterations twice: once using the null.ind:1 setting to test for evidence of admixture accounting for unusual linkage disequilibrium (LD) patterns and once using null.ind:0 to finally infer dates and sources. We further ran 100 bootstraps for the admixture date and calculated the probability of no admixture as the proportion of nonsensical inferred dates (<1 or >400 generations). Confidence intervals were calculated from the bootstraps from the standard model (null.ind:0) using the empirical bootstrap method, and a generation time of 28 years.

ADMIXTURE analysis. We performed ADMIXTURE analysis[41] on the combined Dutch and European samples to explore single marker-based population structure in a set of 41,675 SNPs (LD-pruned using PLINK 1.9: $r^2 > 0.1$; sliding window 50 SNPs advancing 10 SNPs at a time). ADMIXTURE was run for $k = 1$–10 populations, using 5 EM iterations at each k value. The k value with the lowest cross-validation error was selected for further analysis using 15 fold cross-validation; where two k values had equal CV-error the lower k value was taken for parsimony (Supplementary Fig. 10). We analysed the distribution of proportions for each ADMIXTURE cluster across the Dutch dataset, and its relationship with geography.

Computing mean pairwise shared IBD within and between groups. We compared IBD sharing within and between both clusters and provinces (Supplementary Fig. 4) using the mean number of segments within a given length range (e.g. 1–2 cM) shared between individuals. To calculate this mean for a single group of size N with itself the denominator was $(N^2 - N)/2$; when comparing two groups of sizes N and M the denominator was NM.

Estimating recent changes in population sizes. We used IBDNe[26] to estimate historical changes in N_e. IBDNe leverages information from the length distribution of IBD segments to accurately estimate effective population size over recent generations, with a resolution limit of about 50 generations for SNP data. We followed the authors' protocol and detected IBD segments using IBDseq version r1206[42] with default settings and ran IBDNe on the resulting output with default settings, removing IBD segments shorter than 4 cM (minibd = 4, the recommended threshold for genotype data). We compared estimated N_e with recorded census size (https://opendata.cbs.nl/statline/#/CBS/nl/dataset/37296ned/table?ts=1520261958200) for approximately equivalent dates (starting at 1946 CE for generation 0 and assuming 1 generation is 28 years) and found that for generations 0 - 3 our N_e estimates were approximately 1/3 of the census population (Supplementary Fig. 11), which follows expectation if lifespan is ~3× the generation time[26,43]. The slope of the ratios for the three generations is near zero suggesting that our model tracks well with the census population; this is consistent with reported expectation[26].

Estimating effective migration surfaces. To model geographic barriers to geneflow in the Netherlands we ran EEMS[16]. This software provides a visualisation of hot and coldspots for geneflow across a habitat using a geocoded genetic dataset. To run EEMS, we generated an average pairwise genetic dissimilarity matrix from our genotype data using the bed2diffs utility provided with the software. We initially ran the EEMS model with 10 randomly initialised MCMC chains for a short run of 100,000 burn-in and 200,000 sampling iterations, thinning every 999 iterations, to find a suitable starting point. For these runs we placed the data in 800 demes and used default settings with the following adjustments to the proposal variances: qEffctProposalS2 = 0.00008888888; qSeedsProposalS2 = 0.7; mEffctProposalS2 = 0.7. The resulting chain with the highest log-likelihood was then used as the starting point for a further ten chains for 1,000,000 burn-in iterations and 2,000,000 sampling iterations, thinning every 9999 iterations. The model was run with the following adjustments to the proposal variances: qEffctProposalS2 = 0.00008888888; qSeedsProposalS2 = 0.7; mEffctProposalS2 = 0.7. We plotted the results of our analysis using the rEEMSplot package in R and modified the resulting vector graphics using Inkscape v0.91 to remove display artefacts caused by non-overlapping polygons. MCMC convergence was assessed by inspecting the log-posterior traces (Supplementary Fig. 12).

Estimating GWAS confounding. To examine the contribution of observed fine-grained population structure to GWAS confounding, we estimated how well phenotype could be predicted by principal components of haplotype sharing matrices in a 2016 GWAS for ALS[19], comparing our results to those obtained using standard single marker PCA. We separately analysed 1,060,224 zero-missingness Hapmap3 SNPs that passed QC in the original GWAS for Dutch data alone (1963 cases, 2774 controls) and for the complete multi-population GWAS (12,480 cases, 23,275 controls). Haplotypes for unrelated individuals ($\hat{\pi} < 0.075$) were phased using SHAPEIT v2[37] and painted in terms of one another using ChromoPainter v2[20] for the Dutch dataset (estimating N_e and μ using the weighted average of 10 EM iterations on chromosomes 1, 8, 15 and 20 in 10% of samples), and PBWT-paint (https://github.com/richarddurbin/pbwt) for the considerably larger multi-population GWAS dataset. PBWT-paint is a fast approximate implementation of ChromoPainter suitable for large datasets. PCs of the resulting coancestry matrices were calculated using the fineSTRUCTURE R tools (http://www.paintmychromosomes.com), removing extreme haplotype PCA outliers (>20 SD from mean on PC1-10) followed by repainting as an additional QC step. For comparison we also calculated PCs on independent markers from the SNP datasets using Plink v1.9, first removing long range LD regions[44] (https://genome.sph.umich.edu/wiki/Regions_of_high_linkage_disequilibrium_(LD)) and pruning for LD (--indep-pairwise 500 50 0.8). Variance in ALS phenotype explained by ChromoPainter/PBWT-paint PCs and SNP PCs (Nagelkerke R^2) was estimated using the glm() function and fmsb package[45] in R version 3.2.3. To estimate confounding in GWAS inflation, we implemented a logistic regression model GWAS (--logistic) in PLINK v1.9 for each dataset using a range of ChromoPainter/PBWT-paint PCs or SNP PCs (10, 20, 30 and 40 PCs) as covariates and ran LD score regression[46] on the resulting summary statistics using recommended settings (Fig. 6 and Supplementary Fig. 13). Structure evident in the PBWT-paint matrix was visualised and contrasted with corresponding SNP data in 2 dimensions using t-distributed stochastic neighbour embedding (t-SNE)[47] implemented in the Rtsne package in R version 3.2.3 (5000 iterations; perplexity 30; top 100 PCs provided as initial dimensions).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data used in this study are available for academic use through the Project MinE Consortium at https://www.projectmine.com/research/data-sharing/. MS GWAS data used for European reference populations were downloaded from the European Genome-phenome Archive under accession EGAD00000000120. Data availability subject to any conditions outlined by source studies.

Received: 15 January 2020; Accepted: 21 August 2020;
Published online: 11 September 2020

References

1. Abdellaoui, A. et al. Population structure, migration, and diversifying selection in the Netherlands. *Eur. J. Hum. Genet.* **21**, 1277–1285 (2013).
2. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46**, 818–825 (2014).
3. Lawson, D. J. et al. Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? *Hum. Genet.* https://doi.org/10.1007/s00439-019-02014-8 (2019).
4. Leslie, S. et al. The fine-scale genetic structure of the British population. *Nature* **519**, 309–314 (2015).
5. Gilbert, E. et al. The Irish DNA Atlas: revealing fine-scale population structure and history within Ireland. *Sci. Rep.* **7**, 17199 (2017).
6. Byrne, R. P. et al. Insular Celtic population structure and genomic footprints of migration. *PLoS Genet.* **14**, e1007152 (2018).
7. Gilbert, E. et al. The genetic landscape of Scotland and the Isles. *Proc. Natl Acad. Sci. USA* **116**, 19064–19070 (2019).
8. Kerminen, S. et al. Fine-Scale Genetic Structure in Finland. *G3* **7**, 3459–3468 (2017).
9. Takeuchi, F. et al. The fine-scale genetic structure and evolution of the Japanese population. *PLoS One* **12**, e0185487 (2017).
10. Raveane, A. et al. Population structure of modern-day Italians reveals patterns of ancient and archaic ancestries in Southern Europe. *Sci. Adv.* **5**, eaaw3492 (2019).
11. Saint Pierre, A. et al. The genetic history of France. *Eur. J. Hum. Genet.* https://doi.org/10.1038/s41431-020-0584-1 (2020).
12. Bycroft, C. et al. Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. *Nat. Commun.* **10**, 551 (2019).
13. Chacón-Duque, J.-C. et al. Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. *Nat. Commun.* **9**, 5388 (2018).
14. Hellenthal, G. et al. A genetic atlas of human admixture history. *Science* **343**, 747–751 (2014).
15. Novembre, J. & Peter, B. M. Recent advances in the study of fine-scale population structure in humans. *Curr. Opin. Genet. Dev.* **41**, 98–105 (2016).
16. Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48**, 94–100 (2016).
17. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
18. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44**, 243–246 (2012).
19. van Rheenen, W. et al. Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.* **48**, 1043–1048 (2016).
20. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
21. Sawcer, S. et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219 (2011).
22. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).

23. Lawson, D. J. & Falush, D. Population identification using genetic data. *Ann. Rev. Genom. Hum. Genet.* **13**, 337–361 (2012).
24. Palamara, P. F. Population genetics of identity by descent. Preprint at https://arxiv.org/abs/1403.4987 (2014).
25. Al-Asadi, H., Petkova, D., Stephens, M. & Novembre, J. Estimating recent migration and population-size surfaces. *PLoS Genet.* **15**, e1007908 (2019).
26. Browning, S. R. & Browning, B. L. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *Am. J. Hum. Genet.* **97**, 404–418 (2015).
27. Herlihy, D. *The Black Death and the Transformation of the West*. (Harvard University Press, 1997).
28. Durbin, R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinforma. Oxf. Engl.* **30**, 1266–1272 (2014).
29. Athanasiadis, G. et al. Nationwide genomic study in Denmark reveals remarkable population homogeneity. *Genetics* **204**, 711–722 (2016).
30. Nalls, M. A. et al. Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics. *PLoS Genet.* **5**, e1000415 (2009).
31. Roosen, J. & Curtis, D. R. The 'light touch'of the Black Death in the Southern Netherlands: an urban trick? *Econ. Hist. Rev.* **72**, 32–56 (2019).
32. Abdellaoui, A. et al. Association between autozygosity and major depression: stratification due to religious assortment. *Behav. Genet.* **43**, 455–467 (2013).
33. Haworth, S. et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat. Commun.* **10**, 333 (2019).
34. Abdellaoui, A. et al. Genetic correlates of social stratification in Great Britain. *Nat. Hum. Behav.* **3**, 1332–1342 (2019).
35. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
36. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
37. Delaneau, O., Marchini, J. & Zagury, J.-F. cois. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).
38. International HapMap Consortium et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
39. CoreTeam, R. R.: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; (2015).
40. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R. J.* **8**, 289–317 (2016).
41. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
42. Browning, B. L. & Browning, S. R. Detecting identity by descent and estimating genotype error rates in sequence data. *Am. J. Hum. Genet.* **93**, 840–851 (2013).
43. Felsenstein, J. Inbreeding and variance effective numbers in populations with overlapping generations. *Genetics* **68**, 581–597 (1971).
44. Price, A. L. et al. Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.* **83**, 135–139 (2008).
45. Nakazawa, M. *fmsb: functions for medical statistics book with some demographic data, 2014* (R Package, 2018).
46. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
47. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

Acknowledgements

This work has been supported by Science Foundation Ireland (17/CDA/4737), the Motor Neurone Disease Association of England, Wales and Northern Ireland (957-799) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 772376 – EScORIAL). The collaboration project is co-funded by the PPP Allowance made available by Health~Holland, Top Sector Life Sciences & Health, to stimulate public-private partnerships. The authors wish to acknowledge the DJEI/DES/SFI/HEA Irish Centre for High-End Computing (ICHEC) for the provision of computational facilities and support.

Author contributions

R.P.B. and R.L.M. conceived the study. R.P.B., W.V.R., J.H.V. and R.L.M. contributed to study design. R.P.B. and R.L.M. conducted the analyses. R.P.B. and R.L.M. draughted the manuscript. W.V.R., L.H.V.D.B. and J.H.V. provided data and critical revision of the manuscript.

Competing interests

The authors declare no competing interests.

Ethics

Sample collection and data sharing were approved by country-specific institutional review boards and informed consent was obtained from study participants as detailed in the source studies[19,21].

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-18418-4.

Correspondence and requests for materials should be addressed to R.P.B. or R.L.M.

Peer review information *Nature Communications* thanks Abdel Abdellaoui and Javier Mendoza-Revilla for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.