Fine-Scale Estimation of Location of Birth from Genome-Wide Single-Nucleotide Polymorphism Data

Clive J. Hoggart,^{*,1,3} Paul F. O'Reilly,^{*,1} Marika Kaakinen,[†] Weihua Zhang,^{*} John C. Chambers,^{*,‡,§}

Jaspal S. Kooner,^{†,***} Lachlan J. M. Coin,^{*,2} and Marjo-Riitta Jarvelin^{*,†,††,‡‡}

^{*}Department of Epidemiology and Biostatistics, Imperial College, London W2 1PG, United Kingdom, [†]Institute of Health Sciences and Biocenter Oulu, University of Oulu, FIN-90014 Oulu, Finland, [‡]Ealing Hospital National Health Service (NHS) Trust, Middlesex UB1 3HW, United Kingdom, [§]Imperial College Healthcare NHS Trust, London W12 1NY, United Kingdom, ^{***}National Heart and Lung Institute, Imperial College London, Hammersmith Hospital, London W12 0HS, United Kingdom, ^{††}National Institute for Health and Welfare, FI-00271 Helsinki, Finland, and ^{‡‡}Medical Research Council–HPA Centre for Environment and Health, Imperial College, Faculty of Medicine, London W2 1PG, United Kingdom

ABSTRACT Systematic nonrandom mating in populations results in genetic stratification and is predominantly caused by geographic separation, providing the opportunity to infer individuals' birthplace from genetic data. Such inference has been demonstrated for individuals' country of birth, but here we use data from the Northern Finland Birth Cohort 1966 (NFBC1966) to investigate the characteristics of genetic structure within a population and subsequently develop a method for inferring location to a finer scale. Principal component analysis (PCA) shows that while the first PCs are particularly informative for location, there is also location information in the higher-order PCs, but it cannot be captured by a linear model. We introduce a new method, pcLOCATE, which is able to exploit this information to improve the accuracy of location inference. pcLOCATE uses individuals' PC values to estimate the probability of birth in each town and then averages over all towns to give an estimated longitude and latitude of birth using a fully Bayesian model. We apply pcLOCATE to the NFBC1966 data to estimate parental birthplace, testing with successively more PCs and finding the model with the top 23 PCs most accurate, with a median distance of 23 km between the estimated and the true location. pcLOCATE predicts the most recent residence of NFBC1966 individuals to a median distance of 47 km. We also apply pcLOCATE to Indian individuals from the London Life Sciences Prospective Population Study (LOLIPOP) data, and find that birthplace is predicated to a median distance of 54 km from the true location. A method with such accuracy is potentially valuable in population genetics and forensics.

PRINCIPAL component analysis (PCA) has been used extensively to control for population structure (Price *et al.* 2006) and describe genetic diversity. It has been applied to worldwide allele frequency data of human populations (Cavalli-Sforza *et al.* 1993) and more recently to dense SNP data to illustrate population structure in Europe (Lao

et al. 2008; Novembre *et al.* 2008), India (Reich *et al.* 2009), and China (Xu *et al.* 2009). Recent analyses of populations sampled across Europe demonstrated that genetic data could be used to predict place of birth to within a few hundred kilometers (Novembre *et al.* 2008).

Here we introduce a new Bayesian method, pcLOCATE, for predicting location of origin. We apply the method to the Northern Finland Birth Cohort 1966 (NFBC1966), which includes genome-wide data and location information on 2823 unrelated individuals born in 1966 in the two northernmost provinces of Finland, Oulu, and Lapland (Sabatti *et al.* 2008), and to The London Life Sciences Prospective Population Study (LOLIPOP), comprising genome-wide data and location data on 1574 individuals living in West London (UK) and born in India (Chambers *et al.* 2010). We compare

Copyright © 2012 by the Genetics Society of America

doi: 10.1534/genetics.111.135657

Manuscript received October 10, 2011; accepted for publication November 8, 2011

Supporting information is available online at <http://www.genetics.org/content/suppl/2011/11/18/genetics.111.135657.DC1>.

[†]These authors contributed equally to this work.

²Present address: Department of Genomics of Common Disease, School of Public Health, Imperial College London, Hammersmith Hospital, London, United Kingdom.

³Corresponding author: Department of Epidemiology and Biostatistics, Imperial College, St. Mary's Campus, Norfolk Place, London W2 1PG, United Kingdom. E-mail: c.hoggart@imperial.ac.uk

the accuracy of pcLOCATE with the approach of Novembre *et al.* (2008), in which linear models for longitude and latitude were employed with linear and quadratic effects for the first two PCs (Novembre *et al.* 2008). PCA of the NFBC1966 data shows that higher-order PCs appear to contain information on location, but not through a simple linear relationship. pcLOCATE was developed to utilize as many PCs as are informative without assuming a linear relationship between PCs and location. We test the accuracy of the two methods with inclusion of successively more PCs. While the linear method gains little in accuracy with the addition of PCs of order greater than two, pcLOCATE can exploit the information in higher-order PCs and is thus able to achieve finer-scale estimates. By applying pcLOCATE to two population samples with very different characteristics we are able to gain insight into the general applicability of pcLOCATE to human populations. Fine-scale estimation of individuals' location may lead to improved methodology in population genetics and could have immediate application in forensics through assisting in perpetrator identification.

Applying PCA to NFBC1966 also allows us to study the population structure in Northern Finland, one of the most keenly studied population isolates. The settlement of Finland has been characterized by early and late settlement regions (Peltonen *et al.* 1999). The early settlement region comprises the coastal regions of the south and west, which have likely been inhabited for many millennia. The late settlement began in the 16th century with migration into the interior of the country from a small area in the southeast of Finland (Peltonen *et al.* 1999). The late settlement was characterized by the establishment of isolated rural populations by relatively small numbers of founding individuals. Strong founder effects are also evident in India; within Indian ancestral groups there is evidence of excess allele sharing and allele frequency differences between groups are larger than in Europe (Reich *et al.* 2009). Before applying pcLOCATE to the NFBC1966 and LOLIPOP data sets we investigate how the results from PCA applied to the NFBC1966 data compare to what is known about the demographic history of Northern Finland.

Materials and Methods

Genotype data

NFBC1966 aimed to recruit all individuals born in Northern Finland in 1966 and contains genotype data from the Illumina 300K chip for 329,091 SNPs in 4793 individuals. A subset of 61,917 SNPs was selected on which we performed PCA; the SNPs were selected such that call rate >99.5%, minor allele frequency >1%, and Hardy–Weinberg equilibrium P -value >0.005 and thinned to satisfy an LD criteria in which $r^2 < 0.2$ for all pairs of SNPs to prevent overrepresentation of regions of high LD. The PCs generated were used in all analyses. For 2823 individuals we have

records of their parents' town of birth and their most recent town of residence, where town refers to any clustered human settlement. Informed consent from all study subjects was obtained using protocols approved by the Ethical Committee of the Northern Ostrobothnia Hospital District. See Sabatti *et al.* (2008) for further details.

The LOLIPOP study is a population-based cohort study of Indian, Asian, and European white males and females, aged 35–75 years, recruited from West London. Country of birth was recorded with other biomedical information, and blood was taken for genetic analysis. The Illumina Human610 was used to genotype the LOLIPOP individuals. Genotype and location data were available for 1574 individuals from the study that were born in India. Samplewise quality control (QC) included removing duplicates, gender information error, low sample call rate (<95%), related individuals, and ethnic outliers. SNP-wise QC included removing SNPs with low call rate (<95%), low frequency (minor allele frequency < 0.01), and low Hardy–Weinberg equilibrium P -values (<10^{−6}). SNPs were thinned prior to PCA according to their correlation with nearby SNPs, using the pruning option in PLINK (Purcell *et al.* 2007). PCA was calculated with the smartpca program in the EIGENSOFT package (v3.0) (Patterson *et al.* 2006). All participants provided written consent for the genetic studies. The LOLIPOP study is approved by the Ealing and St Mary's Hospitals Research Ethics Committees. See Chambers *et al.* (2010) for further details.

pcLOCATE model

pcLOCATE assumes individuals are from discrete areas, which we refer to as towns, and uses an individual's PCs to estimate the probability of them originating from each of the towns in a country or region. Location is predicted as latitude and longitude by calculating a weighted average of the latitude and longitude of each town, weighted according to the town-of-birth probabilities. Thus the model does not impose a linear relationship between PCs and location but does exploit the local smoothness in PC variation. pcLOCATE uses the top p PCs to estimate origin (either birthplace or most recent residence); the model assumes that the k th PC ($k = 1, \dots, p$) value of an individual originating from town t_j ($j = 1, \dots, m$) is normally distributed with mean μ_{jk} and precision ($= 1/\text{variance}$) λ_{jk} . We use the model to predict parental birthplace, birthplace, and most recent residence. When modeling parental birthplace, we generalize the likelihood such that each individual's PC values contribute two independent observations to the likelihood, one at their father's town of birth and one at their mother's town of birth. In the remainder of this section we describe the model for predicting parental birthplace in detail; the models for birthplace and most recent residence follow straightforwardly from this description, replacing location of the mother's and father's birthplace with the location of the individual's birthplace or most recent residence.

Conditional on parental birthplace and utilizing the first p PCs, the likelihood of individual i 's PC values is given by

$$p(\mathbf{x}_{ip} | \boldsymbol{\mu}_j, \boldsymbol{\lambda}_j, T_{i1} = t_{j1}, T_{i2} = t_{j2}) \\ = \prod_{k=1}^p N(x_{ik} | \mu_{j1k}, \lambda_{j1k}) N(x_{ik} | \mu_{j2k}, \lambda_{j2k}),$$

where $\mathbf{x}_{ip} = (x_{i1}, \dots, x_{ip})$ are the first p PC values of individual i , T_{i1} and T_{i2} are the towns of birth of the parents of i , and N is the normal distribution parameterized by the town- and PC-specific means and precisions given by elements of the matrices $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$.

The location of birth of the parents of a new individual is estimated by Bayesian model averaging in which the latitude and longitude of each town is weighted by the posterior probability of being born in each town. When making a prediction of parents' birthplace, we estimate a single location (assuming one parent) rather than considering all possible pairs of locations for the two parents. Therefore, the expectation of the location of birth of the parents of a new individual z with PC values $\mathbf{z} = (z_1, \dots, z_p)$ is estimated by

$$E(\text{location of birth of parents of } z) \\ = \sum_{j=1}^m L_j p(T_z = t_j | \mathbf{z}, D, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ = \kappa \sum_{j=1}^m L_j p(T_z = t_j | D) \prod_{k=1}^p p(z_k | T_z = t_j, D, \alpha_k, \beta_k), \quad (1)$$

where D is the data (the PCs and towns of birth of parents of all N individuals in the sample), T_z is the town of birth of the parent of z , L_j is the latitude and longitude of town t_j , κ is the normalizing constant, and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of hyperparameters described in Equation 2 below.

The likelihood of observing z 's PCs in each town, required for the product in (1), is obtained by integrating out the unknown town-specific means and precisions. For robustness and to obtain an analytic solution for the likelihoods in (1), we assign μ_{jk} and λ_{jk} their conjugate prior,

$$p(\mu_{jk}, \lambda_{jk}) = N(\mu_{jk} | \mu_0, n_0 \lambda_{jk}) \text{Ga}(\lambda_{jk} | \alpha_k, \beta_k), \quad (2)$$

where Ga is the gamma distribution parameterized by shape parameter α and scale parameter β (mean and variance of distribution: α/β and α/β^2 , respectively). The likelihood terms in the product in equation (1) can then be calculated as

$$p(z_k | T_z = t_j, D, \alpha_k, \beta_k) \\ = \int N(z_k | \mu_{jk}, \lambda_{jk}) \frac{p(\mu_{jk}, \lambda_{jk}) \prod_{i \in S_j} N(x_{ik} | \mu_{jk}, \lambda_{jk})}{p(\mathbf{x}_{S_j})} d\mu_{jk} d\lambda_{jk} \\ = \int N(z_k | \mu_{jk}, \lambda_{jk}) p(\mu_{jk}, \lambda_{jk} | \mathbf{x}_{S_j}, \alpha_k, \beta_k) d\mu_{jk} d\lambda_{jk}, \quad (3)$$

where S_j indicates individuals with a parent born in town t_j and $\mathbf{x}_{S_j}$ are the PC values of these individuals [where both parents of an individual are born in t_j , their PC value contributes two observations to the product in (3)]. Taking $n_0 =$

0 (see below for discussion on choice of prior parameters) and integrating yields

$$= \text{St} \left(z_k | \bar{x}_{jk}, \frac{n_j(\alpha_k + n_j/2)}{(n_j + 1)(\beta_k + (1/2)n_j s_{jk}^2)}, 2\alpha_k + n_j \right) \quad (4)$$

(Bernardo and Smith 1994), where St is Student's distribution with probability density function given by

$$\text{St}(x | \theta, \tau, \gamma) \propto \left[1 + \gamma^{-1} \tau (x - \theta)^2 \right]^{-(\gamma+1)/2}$$

and n_j is the number of parents born in town t_j , $\bar{x}_{jk} = (1/n_j) \sum_{a \in S_j} x_{ak}$ and $n_j s_{jk}^2 = \sum_{a \in S_j} (x_{ak} - \bar{x}_{jk})^2$.

The probability of town of origin prior to observing PCs is assumed to reflect the relative population sizes of the towns and is thus given by a conjugate multinomial Dirichlet model; with an uninformative Dirichlet prior in which all parameters = $\frac{1}{2}$ and the observations are the number of births in each town, this probability is given by

$$p(T_z = t_j | D) = \frac{1/2 + n_j}{m/2 + N}. \quad (5)$$

Therefore, the location of birth of parents of z is estimated by substituting (4) and (5) into (1). By averaging over towns we exploit the smooth variation of the PCs with geography. The most recent residence of individuals in the cohort (recorded at age 31) is estimated by replacing parental birthplace with individual residence in (1) and (2).

The parameters of the prior distribution (2) were determined as follows. The gamma prior on λ_{jk} has mean α_k/β_k , and this was set to the mean precision of PC k across all towns. The degree of shrinkage toward the mean is controlled by α_k ; the larger α_k is, the greater the shrinkage, with each unit increase in α_k having the relative effect of two observations from the prior. We set $\alpha_k = 50$ to balance the requirements of robust estimation and local adaptivity. We compare our results with those achieved by using $\alpha_k = 0$ (no shrinkage) and $\alpha_k = 1000$ (hard shrinkage and thus minimal local adaptivity). Sensitivity analyses showed that prediction was better without shrinking μ_{jk} to the mean PC values, so we assigned the parameter a flat vague prior by setting $n_0 = 0$.

We compare pcLOCATE with the model used in Novembre *et al.* (2008); to implement the method, we find the rotation of PCs 1 and 2 that maximizes the correlation between the rotated vectors and latitude and longitude. The rotated elements are given by

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix},$$

where θ is chosen to maximize

$$f(\theta) = \text{cor}(v_1, \text{longitude}) + \text{cor}(v_2, \text{latitude}).$$

The method of Novembre *et al.* (2008) then fits separate linear regression models for latitude and longitude of

birthplace with linear, quadratic, and interaction effects for the rotated PCs 1 and 2 values,

$$\text{latitude} = \phi_0^{(1)} + \phi_1^{(1)}v_1 + \phi_2^{(1)}v_1^2 + \phi_3^{(1)}v_2 + \phi_4^{(1)}v_2^2 + \phi_5^{(1)}v_1v_2 + \varepsilon^{(1)}$$

$$\text{longitude} = \phi_0^{(2)} + \phi_1^{(2)}v_1 + \phi_2^{(2)}v_1^2 + \phi_3^{(2)}v_2 + \phi_4^{(2)}v_2^2 + \phi_5^{(2)}v_1v_2 + \varepsilon^{(2)},$$

where the ϕ s are the regression coefficients and the ε s are normally distributed random errors. To compare with pcLOCATE when using more than the first two PCs, we extend the model above to also include linear and quadratic effects for PCs 3 and above. We also compare with a model with the same formulation as above but that includes only linear effects for the PCs.

In evaluating all models, overfitting is avoided by out-of-sample prediction whereby each individual's town-of-birth probabilities are calculated with that individual removed from the estimation. Thus, for example, in the evaluation of pcLOCATE, for individual i the estimation of town- and PC-specific μ 's and λ 's is conditional on the PCs and towns of birth of parents of all individuals except i . The distance between two points measured by latitude and longitude was calculated using the Haversine formula (Gellert *et al.* 1989).

Results

Population structure in Northern Finland

We began by inspecting the first two PCs in relation to early and late settlement regions, since those are the most likely to highlight any significant distinction according to settlement region. The NFBC1966 region has been divided into six linguistic/geographic regions: three early settlement regions, North Oulu, South Oulu, and West Lapland; and three late settlement regions, Kainuu and East and Central Lapland (Jakkula *et al.* 2008; Sabatti *et al.* 2008); the regions are marked in Figure 1, B and C. Figure 1A displays a scatter plot of PC 1 against PC 2 values for all individuals whose parents were born in the same region (2136 individuals), colored according to the region of parental birth. The plot shows that individuals from the same region are nearby on the PC1/PC2 plane and that the clockwise order of the regions corresponds to their order geographically. However, the relationship between distance on the PC1/PC2 plane and geographic distance is nonlinear. We also note that the regions do not correspond to distinct clusters; genetic distance, as measured by PCs 1 and 2, appears to be driven by separation-by-distance rather than separation-by-region. The rotation of PCs 1 and 2 found to maximize correlation with latitude and longitude was -26° .

Next we calculated the average intensities of PCs 1 and 2, separately, at each town over all individuals with a parent born in the town; thus, each individual's PC was counted twice, once at their mother's place of birth and once at their father's. Parental birthplace was recorded as the town closest to their location of birth. Figure 1, B and C, illustrates the

results for PC 1 and 2, respectively. Each point represents the average intensity of the PC for that town; the redder the color is, the lower the average PC intensity. Figure 1B shows that PC 1 exhibits a smooth linear southeast/northwest gradient, taking highest values in Kainuu in the southeast of the NFBC1966 region and decreasing with distance from Kainuu. PC 2 also exhibits a smooth linear gradient (Figure 1C) but from southwest to northeast. Both PCs exhibit greater north-south variation in the eastern, late settlement, regions than in the western, early settlement, regions, indicating greater genetic diversity in the late settlement regions, in line with previous studies (Peltonen *et al.* 1999; Jakkula *et al.* 2008). The late settlement region of Kainuu in the southeast is at one end of the PC 1 spectrum, while individuals in East and Central Lapland, also late settlement, are as distant from Kainuu in terms of PC 1 as those from the early settlement regions of North and South Oulu. Furthermore, towns in the early settlement regions of North and South Oulu are close to one another on the first two PCs, while towns in West Lapland, also considered an early settlement region, are relatively distant from Oulu on both PCs. These findings support the notion that the genetic stratification reflects geographic separation rather than separation by region or early/late settlement.

Intensity plots of PCs 3–20 (Supporting Information, Figure S1 and Figure S2) show that these PCs also indicate no early/late settlement distinction but, unlike the first two PCs, do not exhibit a linear relationship with geography. They can, for instance, take values at opposite ends of the PC range at nearby towns. For example, Salla and Kuusamo have mean values of PC 6 in the 5th and 95th percentiles of PC 6 (t -test for difference of PC 6 means, $P < 2 \times 10^{-16}$), respectively, despite being <100 km apart. In fact, Kuusamo is a known internal isolate of Finland, whose present population is thought to trace back to ~ 40 founding families from the 17th century (Varilo *et al.* 2003). Such differentiation in the values of higher-order PCs at nearby towns suggests that higher-order PCs may be informative for fine-scale geographic location and that implementing a non-linear model for PCs may best capture the location information in them.

Application of pcLOCATE to NFBC1966

We first predict parental birthplace and individuals' most recent residence by pcLOCATE without the PCs in the model, thus estimating the town of birth probabilities from the prior alone (Equations 1 and 5), which should reflect the population sizes of each town only. Then, starting with PC 1, we added successively higher-order PCs until the addition of further PCs reduces the predictive accuracy of the model. By adding the first 25 PCs we find that the pcLOCATE model is optimized for these data with the inclusion of the first 23 PCs.

The accuracy of the prediction models is summarized in Figure 2. Figure 2A shows a plot of the number of PCs in the model against the mean and median distance between the

Figure 1 Variation in PCs 1 and 2 in NFBC1966. (A) PC 1 vs. PC 2 values for all individuals whose parents were born in the same region, colored according to parental region of birth. (B and C) Mean intensities of PC 1 (B) and PC 2 (C) in each town with >10 parents from a scale of white (high) to red (low). NFBC1966 regions are marked in B and C: WL, West Lapland; CL, Central Lapland; EL, East Lapland; K, Kainuu; NO, North Oulu; and SO, South Oulu. Four areas, two central and the two northernmost, are not assigned to any NFBC1966 region.

predicted and true birthplace for both parents, and Figure 2B shows that between the predicted and true most recent residence of each individual. The mean and median accuracy of pcLOCATE increases as more PCs are included, to a maximum at 23 PCs for parental birthplace prediction and 8 PCs for most recent residence, while the linear model (Novembre *et al.* 2008) shows negligible improvement in accuracy with more than the first 2 PCs. The mean distance between true and predicted parental birthplace is 47 km for pcLOCATE and 57 km for the linear model, while the median distance is 23 km for pcLOCATE and 45 km for the linear model. The mean distance between true and predicted most recent residence is 67 km for pcLOCATE and 71 km for the linear model, while the median distance is 47 km for pcLOCATE and 59 km for the linear model. Figure 2, C and D, compares the cumulative distribution of the distance between the true and estimated location of the best-fitting models for each method. It shows that pcLOCATE gives substantially better fine-scale prediction than the linear model. For example, pcLOCATE predicts parental birthplace to within 10 km for 37% of individuals' parents and to within 50 km for 68% of individuals' parents, while the linear model predicts 3% to within 10 km and 52% to within 50 km. For most recent residence pcLOCATE predicts 20% of individuals to within 10 km and 50% to within 50 km, while the corresponding proportions for the linear model are 3% and 40%, respectively. When both methods have large errors, the linear model performs slightly better, which can be explained by its predictions being more tightly located around the center of the NFBC66 region, thus limiting the size of the errors. Overall, pcLOCATE gave a more accurate prediction of parental birthplace than the linear model for 80% of parents and a more accurate prediction of most recent residence for 70% of individuals. A *t*-test for difference in mean distance between predicted and actual location of parental birthplace and most recent residence, comparing pcLOCATE and the linear model,

gave $P < 2 \times 10^{-16}$ for both comparisons. Application of the method of Novembre *et al.* (2008) with only linear effects for the PCs gave marginally inferior prediction in comparison with the accuracy achieved by the full model, reducing the mean and median accuracy of the best-fitting model by <1 km.

The six regions of NFBC1966 have markedly different population distributions. Figure 3 summarizes the predictive accuracy of pcLOCATE in the six regions. Figure 3A shows a box plot of the distance between the predicted and true location of parental birthplace stratified by region of birth of the parents (for individuals with both parents born in the same region), while Figure 3B shows the same box plots but for prediction of most recent residence. Despite the differences in population distributions, the accuracy of pcLOCATE is similar in all six regions, which suggests that results from pcLOCATE are robust to heterogeneity in population density. The greater accuracy of the models estimating parental birthplace may be a consequence of increased migration levels since the Second World War. The error distribution of pcLOCATE predictions of parental birthplace and most recent residence are shown in Figure S3, A and B.

Application of pcLOCATE to LOLIPOP

Figure 4 summarizes the variation in PCs 1 and 2 in the LOLIPOP cohort. Figure 4A shows a plot of PC 1 vs. PC 2 values for individuals from the five most populous regions in the study population, accounting for 90% of the sample, and Figure 4, B and C, shows the mean intensities of PC 1 and PC 2 in each region. It is notable from these plots that the relationship between genetic variation and geography is less linear than that observed in NFBC1966.

Figure 5A shows a plot of the number of PCs in the pcLOCATE and linear models against the mean and median distance between the predicted and true birthplace for both parents. Again, the accuracy of pcLOCATE increases as more

Figure 2 Predictive accuracy of pcLOCATE and linear model applied to NFBC1966. (A) Mean (solid lines) and median (dashed lines) predictive accuracy for parental birthplace of pcLOCATE (red) and the linear model (blue) by number of PCs. (B) Mean (solid lines) and median (dashed lines) predictive accuracy for most recent residence of pcLOCATE (red) and the linear model (blue) by number of PCs. (C) Cumulative distribution of the distance between true and estimated parental birthplace of the best-fitting models of pcLOCATE (red) and the linear model (blue). (D) Cumulative distribution of the distance between true and estimated most recent residence of the best-fitting models of pcLOCATE (red) and the linear model (blue).

PCs are included, whereas the accuracy of the linear model improves little with the addition of more PCs. Figure 5B compares the cumulative distribution of the distance between the true and the estimated location of the best-fitting models for each method. Again pcLOCATE gives substantially better fine-scale prediction than the linear model. For example, pcLOCATE predicts birthplace to within 50 km for 47% of individuals, while the linear model predicts only 12% to within 50 km. A t -test for difference in mean distance between predicted and actual birthplace, comparing pcLOCATE and the linear model, gave $P < 2 \times 10^{-16}$. The error distributions of pcLOCATE predictions of birthplace are shown in Figure S4. Application of the method of Novembre *et al.* (2008) with only linear effects for the PCs resulted in inferior prediction in comparison with the accuracy achieved by the full model, reducing the mean accuracy by ~ 5 km.

If pcLOCATE was used in conjunction with the LOLIPOP data to infer birthplace of individuals from the general Indian population, then the prior used in this analysis would need to be replaced with one that was representative of the population density of India. However, in our application the prior distribution reflects the geographic distribution of the study population and so is appropriate for estimating the

location of individuals within this study. It would also be preferable to ascertain a sample of individuals that was representative of the geographic distribution of Indians as it has been shown that PCA can be biased by uneven geographical sampling (Novembre and Stephens 2008; McVean 2009). NFBC1966 is not subject to such bias as it aimed to recruit all individuals born in Northern Finland in 1966 and therefore the sample should reflect the population distribution.

Effect of shrinkage of town-specific precision

Shrinkage of town-specific precision is controlled by α . The effects of varying α on the accuracy of pcLOCATE in the analyses of the NFBC1966 and LOLIPOP data are shown in Figure S5 and Figure S6. These plots show that for the LOLIPOP data shrinking the town-specific precisions, λ_{jk} 's, using either $\alpha = 50$ or $\alpha = 1000$ provides superior accuracy to that achieved by using $\alpha = 0$. However, $\alpha = 50$ and $\alpha = 0$ are optimal for the NFBC1966 data, while hard shrinkage, $\alpha = 1000$, has a detrimental effect on prediction. Therefore, the choice of $\alpha = 50$ achieves both the objectives of robust estimation, which is of importance in the LOLIPOP data, and local adaptivity, which is of importance in the NFBC1966 data.

Figure 3 Box plots of distance between predicted and true location of best-fitting models of pcLOCATE, stratified by parental birthplace. (A) parental birthplace; (B) most recent residence. WL, West Lapland; CL, Central Lapland; EL, East Lapland; K, Kainuu; NO, North Oulu; and SO, South Oulu (for individuals with both parents born in the same region).

Figure 4 Variation in PCs 1 and 2 in the LOLIPOP cohort. (A) PC 1 vs. PC 2 values, colored according to region of birth for individuals from the five most populous regions in the study population, accounting for 90% of the sample. (B and C) Mean intensities of PC 1 (B) and PC 2 (C) in each region, from a scale of white (high) to red (low).

Discussion

Our results suggest that the early and late settlement regions, and six linguistic/geographic regions, used here and elsewhere (Jakkula *et al.* 2008; Sabatti *et al.* 2008) do not correspond to genetic subpopulations. However, we assessed population structure using PCA, where SNPs were selected to be uncorrelated, and we therefore did not exploit information on haplotype variation. Nevertheless, a recent study that investigated variation in LD between individuals from different regions of Finland found that while individuals from the NFBC1966 region exhibited higher LD on average than those from the rest of Finland, there were no discernible distinctions within the NFBC1966 region (Jakkula *et al.* 2008).

The smooth variation in the top two PCs could be explained by the late settlement deriving from migrants from across the early settlement regions, migrating according to geographic location. However, it had been believed that the late settlement derives from a small group of people

in the southeast of the country (Peltonen *et al.* 1999; Norio 2003), which would suggest that the smooth variation in the top two PCs is due to migration since the late settlement period. We cannot infer the direction of migration from PCs (McVean 2009; François *et al.* 2010) and so cannot distinguish between these two hypotheses in the present study. We found that some higher-order PCs take extreme values in some towns, although we cannot make inference on population history from these observations as it has been shown that the patterns of variation in higher-order PCs can be markedly different from the patterns of migration that occurred (Novembre and Stephens 2008). For example, under simulation, even when migration is constant and homogeneous in time and space, higher-order PCs tend to exhibit wave-like patterns of variation with location.

Our model for predicting location of origin, pcLOCATE, is better able to capture the relationship of many PCs with geography than a simple linear model for latitude and longitude with linear and quadratic PC effects (Novembre *et al.* 2008), providing improved predictive accuracy. pcLOCATE is similar

Figure 5 Predictive accuracy of pcLOCATE and the linear model applied to the LOLIPOP data. (A) Mean (solid lines) and median (dashed lines) predictive accuracy for birthplace of pcLOCATE (red) and the linear model (blue) by number of PCs. (B) Cumulative distribution of the distance between true and estimated birthplace of the best-fitting models of pcLOCATE (red) and the linear model (blue).

to the linear discriminant analysis method (Egeland *et al.* 2004) recently employed to differentiate between several rural villages in three European countries (O'Dushlaine *et al.* 2010). Like pcLOCATE, the method uses PCs for prediction and assumes that PC values are normally distributed in each town. However, while O'Dushlaine *et al.* (2010) estimate birthplace as the most probable village using the first three PCs, our estimate is a location in terms of longitude and latitude derived from a weighted average of the posterior probabilities over all towns, which exploits the smooth variation in PC intensity with geography and all of the data. Furthermore, once PCs have been computed, pcLOCATE is fully Bayesian and utilizes as many PCs as are informative. By successively adding higher-order PCs we improve the accuracy of our location estimates, until maximized, demonstrating that many PCs can be informative for location.

In applying pcLOCATE to the NFBC1966 and LOLIPOP data sets we found that the method can predict individual origin to a fine scale. In the NFBC1966 analysis, 68% of individuals had parental birthplace predicted to within 50 km of the true location, while 50% of the sample had most recent residence predicted to within 50 km. The birthplaces of 47% of individuals from the LOLIPOP data were predicted to within 50 km. The results from applying pcLOCATE to population samples with different population densities and sample selection criteria suggest that the method may be widely applicable to human populations. The NFBC1966 study attempted to recruit all births in the region in 1966, whereas the LOLIPOP individuals used in our study are individuals born in India currently residing in West London (UK). F_{st} , a measure of population differentiation, between North and South Oulu and West Lapland within the early settlement region of NFBC1966 is between 0.001 and 0.003 and between Central and East Lapland and Kainuu within the late settlement region is between 0.002 and 0.004 (Jakkula *et al.* 2008). These values are comparable to the genetic differentiation across Europe, where average F_{st} = 0.004 between geographic regions within Europe (Novembre *et al.* 2008). However, F_{st} between early and late settlement regions is as high as 0.006 (Jakkula *et al.* 2008). The average F_{st} between subpopulations within India has been estimated to be ~ 0.011 (Reich *et al.* 2009), thus substantially higher than within Finland and Europe.

Founder effects, genetic isolation, and limited migration in Northern Finland (Peltonen *et al.* 1999; Jakkula *et al.* 2008) are reflected in the accuracy by which both pcLOCATE and the method of Novembre *et al.* (2008) estimate parental birthplace and most recent residence in NFBC1966. The method of Novembre *et al.* (2008) gave a mean distance between true and predicted parental birthplace of 57 km; this compares with an accuracy of a few hundred kilometers when applied to individuals distributed across Europe (Novembre *et al.* 2008) and a mean accuracy of ≈ 280 km when applied to the LOLIPOP data. The disparity between the accuracy of the methods when applied to the NFBC1966

and LOLIPOP data is also dependent on geographic scales over which the two data sets are derived, and therefore the larger the potential errors.

As well as for future applications in population genetics we suggest that our findings could prove valuable for the field of forensics now, particularly in Finland and India. DNA left at a crime scene could be used to estimate the perpetrator's most likely region of origin, providing a lead for an investigation with little other information and no match to a criminal DNA database. In addition, we believe that our method is likely to be applicable to many similar populations around the world, especially those with recent founding effects and/or low migration over many generations. The accuracy achieved in this study may be improved further with the forthcoming availability of sequence data. Sequence data will reveal a huge number of recent rare mutations not present on current SNP genotyping platforms, which are likely to show greater geographic differentiation than older mutations and should therefore increase the accuracy of pcLOCATE.

Acknowledgments

We thank Paula Rantakallio (launch of NFBC1966 and 1986) and Outi Tornwall and Minttu Jussila (DNA biobanking). The authors acknowledge the contribution of the late Academician of Science Leena Peltonen. We thank the participants and research staff who made the study possible. NFBC1966 received financial support from the Academy of Finland [project grants 104781, 120315, 129269, 1114194, Center of Excellence in Complex Disease Genetics and Academy of Finland's Responding to Public Health Challenges Research Programme (SALVE)]; University Hospital Oulu, Biocenter, University of Oulu, Finland (75617); the European Commission (biological, clinical and genetic factors for future risk of cardiovascular diseases, Framework 5 award QLGI-CT-2000-01643); National Heart, Lung, and Blood Institute grant 5R01HL087679-02 through the SNP Typing for Association with Multiple Phenotypes from Existing Epidemiologic Data program (1RL1MH083268-01); National Institutes of Health/National Institute of Mental Health (5R01MH63706:02); European Network for Genetic and Genomic Epidemiology (ENGAGE) project and grant agreement HEALTH-F4-2007-201413; and the Medical Research Council (MRC), United Kingdom (G0500539, G0600705, PrevMetSyn/SALVE). The DNA extractions, sample quality controls, biobank upkeep, and aliquotting were performed in the National Public Health Institute, Biomedicum, Helsinki, Finland and supported financially by the Academy of Finland and Biocentrum Helsinki. The LOLIPOP study is supported by the National Institute for Health Research (NIHR) Comprehensive Biomedical Research Centre Imperial College Healthcare National Health Service Trust, the British Heart Foundation (SP/04/002), the MRC (G0700931), the Wellcome Trust (084723/Z/08/Z), and the NIHR (RP-PG-0407-10371). European Union

grant HEALTH-2007-201550 HyperGenes was to C.H., ENGAGE consortium grant P12892_DFHM was to P.O., and Research Council UK Fellowship was to L.C.

Literature Cited

- Bernardo, J., and A. Smith, 1994 *Bayesian Theory*. Wiley, Chichester, UK.
- Cavalli-Sforza, L., P. Menozzi, and A. Piazza, 1993 Demic expansions and human evolution. *Science* 259: 639–646.
- Chambers, J., J. Zhao, C. Terracciano, C. Bezzina, W. Zhang *et al.*, 2010 Genetic variation in SCN10A influences cardiac conduction. *Nat. Genet.* 42: 149–152.
- Egeland, T., H. Bovelstad, G. Storvik, and A. Salas, 2004 Inferring the most likely geographical origin of mtDNA sequence profiles. *Ann. Hum. Genet.* 68: 461–471.
- François, O., M. Currat, N. Ray, E. Han, L. Excoffier *et al.*, 2010 Principal component analysis under population genetic models of range expansion and admixture. *Mol. Biol. Evol.* 27 (6): 1257–1268.
- Gellert, W., S. Gottwald, M. Hellwich, H. Kustner, and H. Kastner, 1989 *The VNR Concise Encyclopedia of Mathematics*, Ed. 2, Chap 12. Van Nostrand Reinhold, New York.
- Jakkula, E., K. Rehnström, T. Varilo, O. Pietilinen, T. Paunio *et al.*, 2008 The genome-wide patterns of variation expose significant substructure in a founder population. *Am. J. Hum. Genet.* 83(6): 787–794.
- Lao, O., T. Lu, M. Nothnagel, O. Junge, A. Freitag-Wolf *et al.*, 2008 Correlation between genetic and geographic structure in Europe. *Curr. Biol.* 18(16): 1241–1248.
- McVean, G., 2009 A genealogical interpretation of principal components analysis. *PLoS Genet.* 5(10): e1000686.
- Norio, R., 2003 Finnish Disease Heritage I: characteristics, causes, background. *Hum. Genet.* 112(5–6): 441–456.
- Novembre, J., and M. Stephens, 2008 Interpreting principal components analyses of spatial population genetic variation. *Nat. Genet.* 40: 646–649.
- Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A. Boyko *et al.*, 2008 Genes mirror geography within Europe. *Nature* 456: 98–101.
- O'Dushlaine, C., R. McQuillan, M. Weale, D. Crouch, A. Johansson *et al.*, 2010 Genes predict village of origin in rural Europe. *Eur. J. Hum. Genet.* 18(11): 1269–1270.
- Patterson, N., A. Price, and D. Reich, 2006 Population structure and eigenanalysis. *PLoS Genet.* 2: e190.
- Peltonen, L., A. Jalanko, and T. Varilo, 1999 Molecular genetics of the Finnish disease heritage. *Hum. Mol. Genet.* 8(10): 1913–1923.
- Price, A., N. Patterson, R. Plenge, M. Weinblatt, N. Shadick *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38(8): 904–909.
- Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. Ferreira *et al.*, 2007 PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* 81: 559–575.
- Reich, D., K. Thangaraj, N. Patterson, A. Price, and L. Singh, 2009 Reconstructing Indian population history. *Nature* 461: 489–494.
- Sabatti, C., S. Service, A. Hartikainen, A. Pouta, S. Ripatti *et al.*, 2008 Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* 41(1): 35–46.
- Varilo, T., T. Paunio, and A. a. Parker, 2003 The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories. *Hum. Mol. Genet.* 12: 51–59.
- Xu, S., X. Yin, S. Li, W. Jin, H. Lou *et al.*, 2009 Genomic dissection of population substructure of Han Chinese and its implication in association studies. *Am. J. Hum. Genet.* 85: 762–774.

Communicating editor: N. A. Rosenberg

GENETICS

Supporting Information

<http://www.genetics.org/content/suppl/2011/11/18/genetics.111.135657.DC1>

Fine-Scale Estimation of Location of Birth from Genome-Wide Single-Nucleotide Polymorphism Data

**Clive J. Hoggart, Paul F. O'Reilly, Marika Kaakinen, Weihua Zhang, John C. Chambers,
Jaspal S. Kooner, Lachlan J. M. Coin, and Marjo-Riitta Jarvelin**

Figure S1 Mean intensities of PCs3-11. NFBC1966 regions are marked on all plots: WL– WestLapland, CL–Central Lapland, EL– EastLapland, K–Kainuu, NO–North Oulu SO–South Oulu. Four areas, two central and the two northernmost are not assigned to any NFBC1966 region. Only towns with more than ten parents are shown.

Figure S2 Mean intensities of PCs12-20. NFBC1966 regions are marked on all plots: WL– WestLapland, CL–Central Lapland, EL– EastLapland, K–Kainuu, NO–NorthOulu SO–SouthOulu. Four areas, two central and the two northernmost are not assigned to any NFBC1966 region. Only towns with more than ten parents are shown.

Figure S3 Error distributions in the NFBC1966. Distribution of (A) difference between true and estimated parental birthplace and (B) difference between true and estimated most recent residence using the best-fitting pcLOCATE models.

Figure S4 Error distributions in LOLIPOP. Distribution of difference between true and estimated birthplace in the LOLIPOP data using the best-fitting pcLOCATE model.

Figure S5 Cumulative distribution of the distance between true and estimated parental birthplace in the NFBC1966 using the best-fitting pcLOCATE models with varying amounts of shrinkage of within town precision, $\alpha=0, 50, 1000$.

Figure S6 Cumulative distribution of the distance between true and estimated birthplace in the LOLIPOP data using the best-fitting pcLOCATE models with varying amounts of shrinkage of within town precision, $\alpha=0, 50, 1000$.