A reduced representation approach to population genetic analyses provides new insights on human dispersal

Francesca Luca1,2, David Witonsky2, Richard Hudson2,3, Anna Di Rienzo2

1Current address: Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI
2Departments of Human Genetics and 3Ecology & Evolution, The University of Chicago, Chicago, IL

Introduction

Owing to rapidly declining costs, next generation sequencing has become an affordable means to perform surveys of sequence variation on a genome-wide scale. Despite the lower sequencing costs, collecting whole genome sequence data for many individuals is still unaffordable for many laboratories. Surveying a large and representative set of unlinked loci, rather than the entire genome, can provide a valuable alternative for many types of studies, especially if the cost of preparing the sequencing target is low. For example, this is the case if the goal is to estimate demographic parameters. Here, we have developed a reduced representation approach for next generation sequencing that is suitable for population genetic analyses in single individual samples. By applying this approach to 19 individuals from a total of 18 different human populations (Figure 4), we provide new insights on human dispersal.

Objectives

• To implement an inexpensive reduced representation protocol for preparing re-sequencing targets
• To develop the analytical tools necessary for making population genetic inferences
• To test the serial founder model of human dispersal based on re-sequencing data
• To estimate the time of the Out of Africa migration
• To provide a time frame for the colonization of Australia based on large-scale re-sequencing data.

Methods: Reduced representation protocol

Features of the sequencing target prepared by our reduced representation protocol.

Reconstructing past demographic scenarios

Testing the serial founder model of human dispersal on re-sequencing data

Table 1. Estimates of split time from the African founder population (years). Our split time estimate for the migration from the ancestral African population (50,776 years ago [st. dev.: 19,265 years; st. error: 4,673 years]) is consistent with both archeological and molecular data. This estimate assumes a single migration event. However, the archeological record supports an early presence of modern humans in Australia (probably as early as 60,000 years ago), suggesting that an independent migratory flux out of Africa reached Oceania.

Summary of Conclusions

• We developed an effective approach to generate next generation sequencing datasets for population genetic analyses.
• This method is inexpensive and applicable to any species for which a draft or complete reference genome sequence is available
• By applying this approach to human samples, we have tested and validated the serial founder model of human dispersal, for the first time on re-sequencing data.
• Our data support an independent migratory flux out of Africa into Oceania 73,465 years ago (st. dev.: 7,985 years). These are the first genetic data supporting an early presence of modern humans in Australia.

Figure 1. Autosomal regions are evenly represented in the re-sequencing data. Reducing representation protocol.

Figure 2. Distributions of the proportion of reads aligned to restriction fragments in the range 40-100 bp for the samples in this study. The plot shows that our protocol is highly reproducible.

Figure 3. Distribution of %G+C content for the sequencing target, which is very similar to the distribution reported for the entire human genome (Lander et al. 2001)

Figure 4. Map of the approximate geographic locations for the populations sampled in this study.

Figure 5. Nucleotide diversity decreases with distance from a location in Eastern Africa (Addis Ababa). Previous studies showed that microsatellite and haplotype diversities decrease in worldwide population samples as a function of distance from Ethiopia (Li et al. 2008; Ramachandran et al. 2005). These findings were interpreted as evidence for a serial founder model of dispersal of human populations out of Africa. We tested this model for the first time on unascertained sequence variation data. Our estimates return an R^2 of 0.7643, which is very similar to what was previously reported (0.7630).

Figure 6. Heat map of the correlation between nucleotide diversity and geographic distance from locations within Africa; correlation coefficients for each location are color-coded (see legend in the figure). Distances from geographic locations in Southeast Africa are the most strongly correlated with nucleotide diversity. This is in agreement with the location of highest correlation between distance and diversity determined in two studies that sampled a larger set of African populations (Henn et al. 2011; Tishkoff et al. 2009).

Figure 7. Neighbor joining tree calculated using the sequence divergence between individuals; M and F stand for male and female, respectively. Consistent with previous genetic variation studies, the deepest lineages lead to African individuals. The Middle Eastern, European and Gujarati samples form a separate clade. As expected, Native Australians are most closely related to the Melanesian Nasioi and the Native American samples branch together with the Naukan Yupik from North Eastern Siberia.