Università degli Studi di Ferrara

DOTTORATO DI RICERCA IN
BIOLOGIA EVOLUZIONISTICA E AMBIENTALE

CICLO XXVIII

COORDINATORE Prof. Guido Barbujani

Lost worlds: tales of archaic hominin admixture in Southeast Asia and Oceania

Settore Scientifico Disciplinare BIO/18

Dottorando	**Tutore**
Dott. Tucci Serena	Prof. Barbujani Guido

Anni 2013/2015

Table of contents 3

List of publications 7

Abbreviations 8

Preface 9

BACKGROUND AND INTRODUCTION 13

Lost worlds: hominin evolution in Island Southeast Asia. **13**

First-come, first-served. H.erectus dispersal in Indonesia 13

Little people from Flores 15

The Flores pygmies 17

Modern humans voyaging in Southeast Asia and Oceania 18

Archaic admixture in the genomic era **18**

Archaic admixture inference 21

Methods relying on ancient DNA data 21

"Fossil free" methods for archaic admixture inference 24

Archaic legacy 28

Research Aims **29**

METHODS 30

Samples collection **30**

Selecting individuals for whole-genome sequencing **30**

Genome-Wide SNP Genotyping 30

Identification of duplicated samples or cryptic relatedness 31

Deconvoluting population affinities in Asia 32

Analysis of whole-genome sequencing data 33

Whole-Genome Sequencing and Filtering *33*

Inference of haplotype phase *34*

Integration of WGS with the Human Origin dataset *35*

Inference of population structure *36*

Estimates of Inbreeding *37*

PSMC analysis *37*

Inference of Denisovan and Neandertal ancestry in the Flores pygmies *38*

Statistical method to identify and classify archaic sequences *39*

Step 1: Calculating S 39

Step 2: Calculating archaic match p-values 40

Step 3: Calling and classifying archaic sequence 41

Estimating ages of unknown introgressed sequences *42*

RESULTS 43

Selecting individuals for whole-genome sequencing 43

Genotype quality and cryptic relatedness *43*

Flores in the context of Asian genetic diversity *46*

The Flores genomes 48

Flores in the context of worldwide genetic diversity *49*

Inbreeding in the Flores pygmies *52*

Inference of long-term effective population sizes *53*

Neandertal and Denisovan gene flow in the Flores pygmies *55*

Ancient introgression in the Flores genomes *59*

DISCUSSION 63

FURTHER APPLICATION OF THE S* STATISTICAL FRAMEWORK 66

Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals 66

Description of the Melanesian samples *67*

Island Melanesians in the context of worldwide population diversity *68*

Melanesians harbour traces of gene flow from Neandertal and Denisova *68*

Identifying Neandertal and Denisovan introgressed sequences........ 69

Neandertal admixture occurred multiple times in different non-African populations 70

Archaic deserts 71

Signatures of adaptive introgression in the Melanesian genomes 71

CONCLUSION AND FUTURE PERSPECTIVES 77

SUPPLEMENTARY TABLES S1-S8........ 78

SUPPLEMENTARY FIGURES S1-S3 89

BIBLIOGRAPHY 92

List of publications

Benjamin Vernot, **Serena Tucci**, Janet Kelso, Joshua G. Schraiber, Aaron B. Wolf, Rachel M. Gittelman, Michael Dannemann, Steffi Grote, Rajiv C. McCoy, Heather Norton, Laura B. Scheinfeldt, David A. Merriwether, George Koki, Jonathan S. Friedlaender, Jon Wakefield, Svante Pääbo and Joshua M. Akey, *Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals*. Science (2016)
Available at http://science.sciencemag.org/content/early/2016/03/16/science.aad9416.abstract
Reprinted with permission from AAAS. License Number 3832910927371

Irene Lobon*, **Serena Tucci***, Marc de Manuel, Silvia Ghirotto, Andrea Benazzo, Javier Prado-Martinez, Belen Lorente-Galdos, Kiwoong Nam, Marc Dabad , Jessica Hernandez-Rodriguez, David Comas, Arcadi Navarro, Mikkel H. Schierup, Aida M. Andres, Guido Barbujani, Christina Hvilsom, Tomas Marques-Bonet. *Demographic history of the genus Pan inferred from whole mitochondrial genome reconstructions (Submitted)*

*These authors contribuited equally to this work

Abbreviations

AMH	Anatomically Modern Humans
ARG	Ancestral Recombination Graph
bp	Base pairs
DNA	Deoxyribonucleic acid
FDR	False discovery rate
HO	Human Origin
IBD	Identity by descent
IBS	Identity by state
ISEA	Island Southeast Asia
Kya	Thousand years ago
LD	Linkage Disequilibrium
MAF	Minor allele frequency
Mya	Milion years ago
PCA	Principal Component Analysis
PSMC	Pairwise Sequentially Markovian Coalescent
QC	Quality Control
ROH	Runs of Homozygosity
SNP	Single Nucleotide Polymorphism
TMRCA	Time to the Most Recent Common Ancestor
VCF	Variant Call Format
Ya	Years ago

Preface

When in 2003 Michael Morwood first uncovered the remains of the fossil now known as *Homo floresiensis* in the Liang Bua cave, he already knew he was treading on dangerous ground, because the Liang Bua fossil would have challenged fundamental beliefs about our recent evolutionary past.

Indeed, a decade on from their discovery, the remains of the "Hobbit" - a small brained, small-bodied hominin from the Indonesian Island of Flores - are still cause for much debate.

Do these remains belong to a new human species, maybe product of insular dwarfism from a larger bodied ancestor? Or, is the "Hobbit" a pathological human suffering of microcephaly? Finally, did these archaic humans interact with modern humans in the area? Did this interaction include events of admixture?

To answer to all these questions I travelled to the remote island of Flores in Eastern Indonesia, where the fossil was found – and where today a pygmy population still lives in a village close to the Liang Bua cave. The team of the expedition included my co-supervisor at UCSC prof. Richard Edward Green, expert in archaic admixture and author of the first Neandertal genome, Pradiptajati Kusuma and Gludhug Aiyopurnomo, students at the Eijkman Institute of Jakarta (our Indonesian counterpart in this project) - and me.

Much as Wallace in 1859, every voyager travelling to Indonesia would be fascinated by the rich diversity of cultures, people and languages for which this corner of the world is home to. Indonesia, the world's largest archipelago, is a chain of more than 17,000 islands that stretches between the continents of Asia and Australia, dividing the Pacific and Indian Oceans.

The almost 240 million Indonesia's inhabitants are extremely diverse, speaking more than 750 languages and representing more than 300 different ethnic groups. Importantly, the Eastern part of Indonesia, hosts the highest linguistic diversity of any Southeast Asian region, as it is the boundary of two completely unrelated linguistic groups, the Austronesian and Papuan linguistic families.

For sure far from the adventurous Wallace's travels, my expedition to Eastern Indonesia starts in Jakarta, where our team spent few days to set up all the logistics for our trip to Flores with the invaluable help of our local collaborators at the Eijkman Institute.

After flying from Jakarta to Bali, and then from Bali to Labuan Bajo, we finally arrived in the Island of Flores. The small rugged Island of Flores is part of the Indonesian province of East Nusa Tenggara, in the eastern part of the Lesser Sunda Islands.

Geographically, Flores is located exactly halfway between the Asian and the Australian continental areas - right on the cultural and linguistic boundary between Asia and Oceania, and strikingly - the island is located on a plausible route for human migrations into Australia.

Heading to Rampasasa, our first stop was Ruteng, the capital of the Manggarai district and strategic center where we could made final arrangements before our visit to Rampasasa.

The winding way from Labuan Bajo Airport to Ruteng by jeep - an authentic nightmare for people suffering car sickness *like me* - has soon revealed us breathtaking landascape views beside the way to Ruteng.

The Manggarai people, who have been living in the area for centuries, make their characteristic spider web ricefields. These series of concentric circles ricefields are sacred to this people who believe that the central point is the site where the sacred spirits reside.

It took us a four hour to get to Ruteng. After getting the necessary permits from the local authorities of the Manggarai regency, and after hiring a field assistant to help us communicate with the villagers speaking Manggarai languages - we were ready to meet with the Rampasasa pygmies. About one hour by jeep from Ruteng on a dirt road through the forest, there is the Rampasasa village.

Getting into a village is exactly the same as getting in someone's house: you first ask to the inhabitants if you are welcome, and wait for their response. And this is what we did. All the villagers were gathered in the main shack of the village, and we were finally asked to join them.

As soon as we got in, it was not hard to notice the large number of very short people, especially among the elders. Hidden in the smoke of cloves cigarettes, the silence was broken by the words of the headman of the village who finally pronunced his approval about our visit. It looked like we had got the benevolence of the living villagers - but for the benevolence of the dead, we still had to wait that a chicken was sacrified to the ancestors on a ritual rock. Despite the fact that the Rampasasa people are Roman Catholics, traditional beliefs about ancestral spirits (*adat*) are still apparent today in their everyday life. Once ancestors were satisfied, we were offered palm wine (*moki*) and invited to drink it together with the elders. We were now asked the reason of our visit. We started explaining very carefully the purpose of our project, and clearly stated our hypothesis concerning the ancestry of the Rampasasa villagers. The more we talked, the more enthusiasm we got from the whole village. We then learned that for many generations – far before the discovery of *H. floresiensis* - the Rampasasa villagers thought their history to be somehow related to the Liang Bua cave. They believe in fact that their ancestors lived in the cave and might have been buried there - even today it is common to find food offers placed at the rear of the cave for the ancestral spirits, and excavations in the younger sedimentary unit at Liang Bua have revealed evidence of modern human burials and a high density of artifacts. Also, since their childhood, the Rampasasa villagers have heard stories of little and hairy people who lived at Liang Bua, and they say it was not uncommon to see them in the forest until very recently: they call the small folk *paju*. These stories are quite common all over Flores, though (i.e. the Nage *ebu gogo* legend).

While the cultural traits of the Rampasasa people have captured our imagination, their phenotipic traits looked even more striking to us.

At the end of our meeting, it was clear to us that the Rampasasa pygmies were definitely interested in becoming involved in our research and shared with us the desire to know more about their evolutionary past.

BACKGROUND AND INTRODUCTION

Lost worlds: hominin evolution in Island Southeast Asia.

Island Southeast Asia (ISEA) plays a key role in the understanding of hominin evolution and migrations, as it was the endpoint of the *Homo erectus* out of Africa dispersal, and likely the waypoint of the modern human exodus resulting in the colonization of New Guinea and Australia at least 50 kya years ago (Roberts et al., 1990; Clarkson et al., 2015; O'Connor, 2015).

Glacio-eustatic sea-levels lowering, documented during Pleistocene (Murray-Wallace, 2014), exposed vast dry lands of the *Sunda Shelf*, a now mostly submerged extension of the Eurasian continental plate, stretching towards Java and the Indian Ocean (Figure 1), creating temporary connections that linked the western islands of today's Indonesian Archipelago (including the island of Java and Borneo), with the Southeast Asian mainland (also known as *Sundaland*).

East of Sunda, the Wallace's line (Huxley, 1868), one of the world's biggest biogeographic disjunctions, marks the limit of the Sunda Shelf and of the Eurasian ecosystems dominated by placental mammals.

Pleistocene interglacial high sea levels contribuited to the isolation of early archaic hominin groups across ISEA, limiting their dispersal into the *Sahul* (the ancient continent currently represented by Australia and its continental islands), setting up conditions for insular endemisms and for their interactions with modern humans populations moving out of Africa and through Asia.

First-come, first-served. H.erectus dispersal in Indonesia.

Indonesia, a remnant of the Sunda ancient continent, has witnessed the evolution of *Homo erectus* lineages for over 2 millions years (Kaifu et al., 2008; Swisher et al., 1994; Swisher et al., 1996), and is home today to a rich fossil documentation of hominin presence.

Any introduction on the Indonesian fossil record should begin with Eugene Dubois who discovered the first Pleistocene hominin fossil in the region. Originally named *Pithecanthropus erectus* — now commonly accepted as *H. erectus* — this human form was first discovered at Trinil in the Solo Basin in eastern Java, by Dubois in 1891 (Dubois, 1894). Since then, the Solo Basin has produced more than 80 *H. erectus* cranial and dental fossils recovered from the sites of Trinil, Sangiran and Ngandong. The face of *H. erectus* protrudes and presents a noticeable large ridge of

bone above the eyes. These fossils show thick cranial vaults and cranial capacities ranging from 813-1251 cc (Baba et al., 2003).

H. erectus has been documented throughout much of the Pleistocene of Java (Swisher et al., 1994; Swisher et al., 1996; Larick et al., 2001), with most of earlier evidence coming from Sangiran, whose specimens are dated to about 1.6–1.7 My ago (Swisher et al., 1994, Larick et al., 2001), while the speciments in Ngandong dated 130-102 kya (Bartstra et al., 1988) or even more recently 53-23 kya (Swisher et al., 1996), may record the last appearance of the lineage in the region (Indriati et al., 2011).

Although its chronological span and phylogenetic origins remain obscure, *H. erectus* lineage is thought to have evolved in Africa before 1.89 million years ago (Walker and Leakey, 1993; Spoor et al., 2007) and have dispersed from there into Asia shortly after (Swisher et al., 1994), becoming the oldest hominin to have ventured out of Africa (Semah et al., 2000).

Lately, this model has been challenged by discoveries at Dmanisi (Republic of Georgia, Southwest Eurasia), where recent excavations uncovered hominin speciments dated to about 1.85 Mya (Ferring et al., 2011), providing evidence for the oldest known occupation of Eurasia. According to Margvelashvili et al. (2013) and Lordkipanidze et al. (2013), the wide phenotypic diversity observed among specimens of the same time period should lead us to reconsider major aspects of the current hominin classification, which I shall not address in this thesis, maintaining the traditional *H.erectus* definition. Indeed, Dmanisi's rich collection, which now comprises five crania, is related to *H. erectus*. Interestingly, the Dmanisi's sample fossils, which revealed a population with cranial capacities of only 600–775 cc and post-cranial skeletons shorter compared to *H. erectus*, exhibits close morphological affinities with the earliest known *Homo* fossils from Africa (Lordkipanidze et al., 2013). Following these observations, several authors suggested that the Dmanisi hominids might be descendants of *H. habilis*–like ancestors, who dispersed from Africa before *H. erectus* (Vekua et al., 2002).

Taken together, evidences from the Dmanisi site and from the Sangiran dome, suggest that ISEA and Southwest Eurasia have been theaters for one of the major evolutionary shifts in human evolution, the dispersal of *Homo* out of Africa, potentially even earlier than previously thought, with clear implications for our understanding of taxonomic diversity and interpretations of the broad geographic and temporal distribution of hominin speciments.

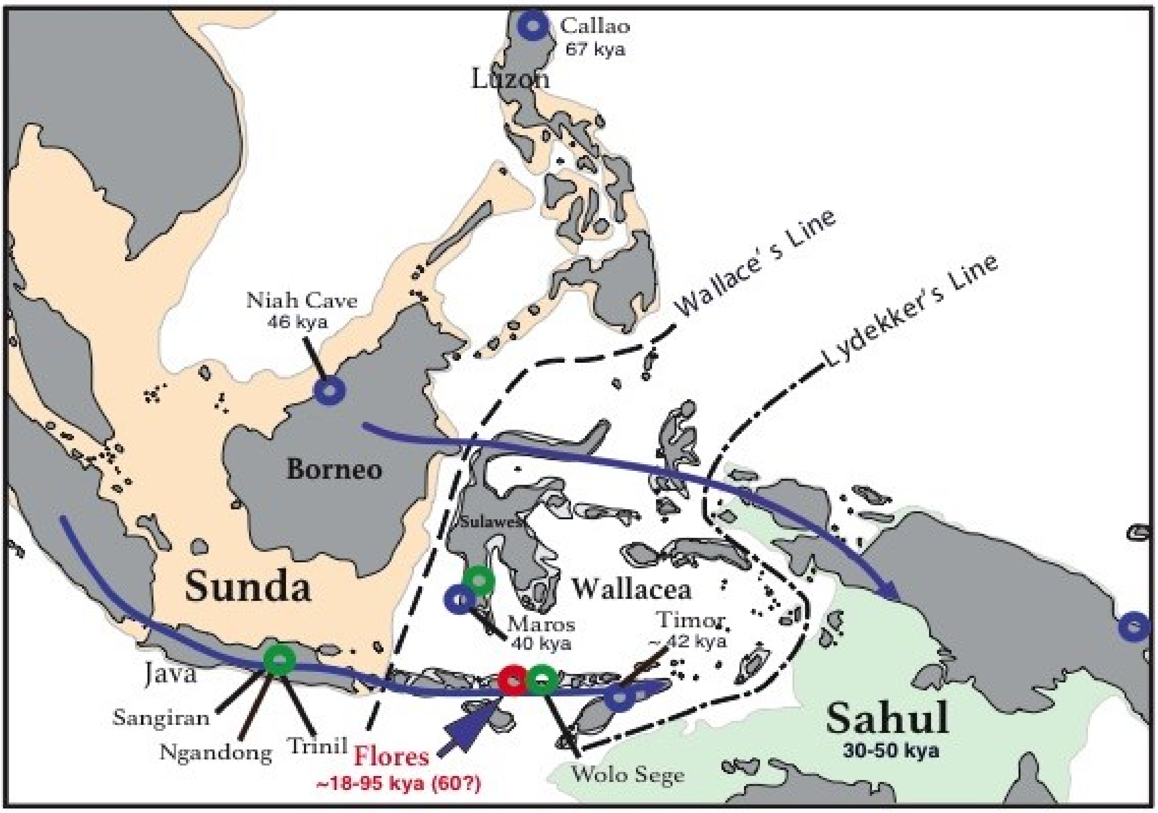

Figure 1. Archaeological evidences for hominin presence in Southeast Asia. Red circle indicates the approximate location of Liang Bua cave (Flores). Circles show early hominin sites (in green) and modern human sites (in blue). Wallace's Line marks the limit of the Sunda shelf (orange) and of the Eurasian placental mammals distribution. Lydekker's Line represent the border of the Sahul shelf (green) and limits marsupial mammals distribution. The area between the two lines is known as "Wallacea".
Blue lines represent hypothetical Northern and Southern coastal routes of modern human dispersal into the Sahul.

Little people from Flores

In October 2003, the Indonesian fossil hominin record enriched by the addition of a new hominin species. Found in the limestone cave of Liang Bua on the island of Flores (Eastern Indonesia) (Brown et al., 2004; Morwood et al., 2004) – *Homo floresiensis* has attracted since then a worldwide interest, virtually unparalleled in the annals of palaeoanthropology.

The hominin fossils now assigned to *H. floresiensis* consisted of a partial skeleton "LB1" (that later became the type specimen of *Homo floresiensis*), plus the remains of several other individuals (Brown et al., 2004; Morwood et al., 2005; Morwood et al., 2009) recovered from the Late Pleistocene deposits and dated between ~18 and 74-95 kya (Brown et al., 2004; Morwood et al., 2004), although a new minimal age of ~60 kya has been recently proposed (Larick and Ciochon 2015, from personal communication of A. Brumm).

Described as a small-bodied bipedal hominin with endocranial volume of 417 cm^3 (Falk et al., 2005), similar to that of *Australopithecus afarensis* specimen AL 288-1 (Lucy) who lived approximately 3 million years ago - *H. floresiensis* has been interpreted in widly different ways since its discovery. Several skeptics have initially argued that *H. floresiensis* was a pathological modern human suffering from growth anomalies, microcephaly, Laron syndrome or cretinism (Jacob et al., 2006; Henneberg and Thorne, 2004, Oxnard et al., 2010). However several studies eventually showed that methodologies used by those opponents of the new species argument (Weber et al., 2005; Jacob et al., 2006; Oxnard et al., 2010) were not appropriate, and/or data were not collected/reported properly (Van Heteren, 2013). In addition, none of the diseases proposed so far was shown to be compatible with the entire suite of features evident in the fossils. For those reasons, the pathological argument currently seems to have been dismissed for good (Falk et al., 2005; 2007, 2009, 2010; Jungers et al., 2009; Larson et al., 2007). The consensus now appears to be that the individuals recovered from Liang Bua cave retain a mosaic of both derived and primitive traits, indicating that they may be insular dwarf descendants of a *H. erectus* population (Brown et al., 2004; Falk et al., 2005; Gordon et al., 2008; Kaifu et al., 2011) - or of an even earlier - *pre-erectus* hominin species, whose dispersal into Southeast Asia is still undocumented (Morwood et al., 2005; Argue et al., 2009; Morwood et al., 2009; Jungers et al., 2009).

Notably, a *H. erectus* ancestry better accords with the current biogeographical evidence of a well known presence of *H. erectus* in nearby Java, but implies a more marked body size reduction than an earlier origin. Recent evidence from the study of fossils hippo that underwent insular dwarfing in Madagascar, showed that natural selection can act shrinking brains to volumes well below the sizes predicted by the scaling trend (Weston and Lister, 2009). This study provided significant support to the hypothesis that insular dwarfism might have played a role in the evolution of *H. floresiensis*, likely from a hominin species presents in Flores as early 1 Myr ago, as documented by the Wolo Sege stone artefacts, which currently comprise the oldest evidence for hominins on Flores (Brumm et al., 2010).

The Flores pygmies

The unique set of morphological features of the Liang Bua hominins has inspired a great deal of research and commentary for over ten years. Among the opponents of the view whereby a new species had been discovered, Jacob et al. (2006) is important to the present project. Indeed, Teuti Jacok belief that *H. floresiensis* represented a pathological modern human and not a new species, led his team to travel to the island of Flores, where he discovered a community of individuals who are small in stature – 80% were small enough to be considered Pygmies. Notably, this population lives in the Rampasasa village, in close proximity of the Liang Bua cave in which *H. floresiensis* fossils were uncovered.

Modern human pygmy populations are widely distributed globally, and researchers have shown that their short stature appears to represent one aspect of a complex eco-geographic adaptation to rain forest or island environments (Perry, 2009). Although numerous genetic studies have been conducted on pygmy populations, such as the Baka in Africa (Batini et al., 2011; Jarvis et al., 2012; Lachance et al., 2012; Perry et al., 2014) or the Negritos in Southeast Asia (Reich et al., 2011; Jinam et al., 2013; Aghakhanian et al., 2015), nothing is known about the genetic history of the Indonesian pygmies of Rampasasa. The people of this village are of acute interest because of their unique phenotypic features associated with the region they inhabit. While the cranial volumes differ considerably from *H. floresiensis*, Jacob et al (2006) showed that *H. floresiensis* do share features (receding chins and rotated premolars) with the Rampasasa pygmies. Moreover, local folklore claims that these individuals are admixed descendants of *H. floresiensis*. On the light of those observations, Jacob suggested that Liang Bua fossils exhibit a combination of regional characters with no substantial deviation compared to the Rampasasa pygmies. Therefore, while presenting signs of a developmental abnormality, including microcephaly – he concluded that the fossils belong to an earlier modern human pygmy (Jacob et al., 2006).

Although there is no doubt that the pathological argument for *H. floresiensis* is not supported by any evidence, most of the opposition to the new species explanation comes from the fact that the Liang Bua discovery has forced paradigm-changing research in several areas of anthropology (Aiello, 2010). It has also raised a number of questions, one of them concerning the possible interaction of such archaic humans with modern humans in Island Southeast Asia.

Modern humans voyaging in Southeast Asia and Oceania

Archaeological evidences suggest that modern humans arrived in Australia as early as 50 kya (Roberts et al., 1990; Clarkson et al., 2015), in Papua New Guinea between 50-30 kya (Groube et al., 1986; Wickler and Spriggs, 1988; Leavesley and Chappell, 2004), and in Borneo around 46 kya (Barker et al., 2007). Then, Island Southeast Asia – and in particular Eastern Indonesia - was likely the crossroad of multiple migrations from Sunda to Sahul

(Figure 1), but the routes and timing of those dispersal in the area remain unclear. The expansion from Sunda to Sahul might have occurred along a Northern route (via Sulawesi to Papuan New Guinea), or a Southern "Wallacea" route, via Nusa Tenggara (from Sumatra, Java, Lombok, Flores and Timor). In the last case, upon reaching what is now Java and Borneo, the presence of deep-water trenches of Wallacea would have impeded further migrations of these early voyagers.

However the discovery of the archaelogical site of Jerimalia in East Timor, dated to 42 kya (O'Connor, 2007), provided evidence that the Southern route to Sahul, through Nusa Tenggara – and notably Flores – was a reliable alternative to the Northern route, for modern humans able to perform some sort of seafaring. Then, the confirmation of Paleolithic settlers in the region, suggested that the initial colonization of Eastern Indonesia by modern human might have occurred more than 50,000 ya (Spriggs, 2000), overlapping significantly in time with *H. floresiensis* in the region.

Archaic admixture in the genomic era

The question of modern human origins has fascinated anthropologists for decades and has been the object of heated scientific debates. We now know, by the study of the fossil record, that anatomically modern humans (AMH) evolved in Northeast Africa around 200,000 years ago (McDougall et al., 2005; Fleagle et al., 2008). These early humans quickly expanded across Africa, and dispersed from there between 120,000 and 60,000 years ago, eventually replacing all other existing hominin forms (Ramachandran et al., 2005; Liu et al., 2006; Stewart et al., 2012). The exact number of major dispersals from Africa, and their timing, are still poorly defined, but both archaeological (Armitage et al. 2011) and genomic data (Tassi et al. 2015) suggest that an early migration episode may have led to the peopling of Southeast Asia through a Southern route, crossing the Arab peninsula and the Indian subcontinent.

In the course of these expansions, anatomically modern groups encountered populations of archaic morphology, in Europe and Asia. Still debated are the genetic consequences of the interaction between these two groups, including Neandertals in Europe, Denisovans and perhaps other human forms in Asia. In particular, the extent to which anatomically modern humans may have admixed with archaic humans, and what contribution these archaic human populations might have made to the contemporary human gene pool, is still unclear.

For decades, these questions remained in the domain of anthropologists who study the morphological features of fossil bones. Those features, taken together, can diagnose distinct human species in the past. However, morphological features of present-day humans and early AMH fossils have been interpreted as evidence both for and against interbreeding between Neandertals and AMH (Trinkaus et al., 2003, Brauer et al., 2006).

That some episodes of interbreeding have happened seems now out of discussion (Fu et al., 2015; Kuhlwilm et al., 2016); the question is whether a hybrid population derived from such crosses, and whether it has left descendants up to the present time.

Today, DNA sequences retrieved from hominin bones offer a new approach for understanding hominin relationships and have enabled these long-standing questions to be addressed directly.

Indeed, the development of high-throughput DNA sequencing technologies has allowed for the genome-wide sequencing of nuclear DNA from ancient specimens, culminating in the first draft sequence of the Neandertal genome in 2010 (Green et al., 2010). These data, along with DNA data from several living humans, show that populations of AMH differ in their degree of genetic similarity to Neadertals. All studied genomes from Europe, Asia and Papua New Guinea appear slightly but consistently more similar to Neandertal genomes than the African genomes. Although at odds with those obtained from mtDNA (Currat and Excoffier, 2004; Serre et al., 2004; Ghirotto et al., 2011), these results have been interpreted as evidence of admixture between Neandertals and human ancestors – probably soon after leaving Africa – to an extent that, on average, 1–4% of the genomes of people outside Africa might derived from Neanderthals (Green et al., 2010).

Incidentally, the commonly accepted hypothesis that could explain this observation is a history of gene flow from Neandertals into early modern humans, presumably when they encountered each other in Europe and the Middle East. However, an alternative possibility exists, and is related with the existence of population structuring in ancient Africa. That way, one would expect a closer similarity between Neandertals and modern humans outside Africa, than between Neandertals and modern humans from sub-Saharan Africa. If this substructure persisted until modern humans expanded out of Africa, it is not necessary to assume hybridization between Neandertals and modern humans to account for the fact that today's people outside Africa share more genetic variants with Neandertals than people in sub-Saharan Africa.

The sequencing of another previously undescribed archaic hominin from the Denisova cave, in the Altai mountains (Siberia), has further revealed that Papua New Guineans harbor signs of admixture with this archaic human (Reich et al., 2010). Recent studies of several populations from Southeast Asia and Oceania showed that Denisovan admixture is actually present in other Islands East of the Wallace Line in Oceania (Reich et al., 2011; Skoglund and Jakobsson, 2011; Qin and Stoneking 2015).

Although caution is urged in inferring recent admixture between ancient hominins and modern-day populations, recent research suggests that hybridization is not uncommon in the wild today between closely related species, including primates, either by forming hybrid zones, or hybridizing during periods of environmental disturbance, or if the species are rare (Zinner et al., 2011; Jolly, 2011).

Since the publication of the first Neandertal and the Denisova genomes, the issue of gene flow between modern humans and archaic forms, received considerable attention. Several groups have started to analyze DNA sequence variation from living human populations with the goal of identifying regions of archaic ancestry and several tests of admixture have been devised, mainly relying on the pattern of shared polymorphisms between modern humans and ancient genomes, such as the well known Patterson's *D statistics* (Green et al., 2010; Patterson et al., 2012).

Hence, using similar approaches, gene flow between AMH and archaic forms has been recently described for several populations worldwide (Meyer et al., 2012; Wall et al., 2013; Lazaridis et al., 2014; Prufer et al., 2014; Qin and Stoneking, 2015).

Whenever the genome of a candidate archaic contributor to admixture is known, the rationale of the tests is simple. The null hypothesis is that two modern genomes share the same proportion of derived alleles with an archaic genome; the alternative hypothesis is that an excess similarity of one modern genome reflects admixture with the archaic form (Green et al. 2010; Durand et al. 2011; Martin et al. 2015). However, our knowledge of extinct human groups is very limited, and even more so our knowledge of their genomes. Therefore, the question arises whether it is possible to detect evidence of archaic admixture, even in the absent of genetic information about the archaic population.

To address this question, another class of methods for archaic admixture inference has been developed, relying on the signature that ancient admixture leaves on patterns of linkage disequilibrium (LD) in the genomes of present-day humans. Among that category, a new computational approach, also known as "S* statistic" - has been recently developed and applied to analyse whole-genome sequencing data from several contemporary human populations (Lachance et al., 2012; Vernot and Akey, 2014; Hsieh et al., 2016). Notably, while providing further evidence of admixture with archaic human species, this method showed that - by analyzing modern human genomes, significant amounts of DNA sequences of these archaic groups can be retrieved - even in the *absence* of fossilized remains from the archaic species (Lachance et al., 2012; Vernot and Akey, 2014; Hsieh et al., 2016).

Archaic admixture inference

Methods relying on ancient DNA data

Green et al 2010, introduced a formal test for admixture, based on the direct comparison of DNA sequences from archaic and modern human populations, to evaluate whether modern humans differed in their genetic similarity to Neandertals.

This new statistic, defined as *D statistic* (Equation 1), tests whether a given phylogenetic tree applies to the data. The statistic is based on an ordered set of populations (*P1, P2, P3, O*), and focus on biallelic sites at which an allele ("A") is seen in the *outgroup O*, the alternative allele "B" is seen in *P3*, and the population tested *P1* and *P2* differ:

$$D(P1, P2, P3, O) = \frac{C_{ABBA(i)} - C_{BABA(i)}}{C_{ABBA(i)} + C_{BABA(i)}} \qquad (1)$$

$$E(D(P1, P2, P3, O) = 0$$

Thus, the two possible patterns of single nucleaotide polymorphism (SNP), termed "ABBA" or "BABA" are observed at site *i* in the genome (Green et al., 2010). Assuming that the SNPs considered in the computation are ascertained as polymorphic in an outgroup O, and if the tree assumption is correct, under the null hypothesis of no gene flow, we expect an equal rate of ABBA and BABA sites, and then, the numerator of *D* to be close to zero. The asymmetry in the three population gene trees is then interpreted as evidence for gene flow between the populations *P2* and *P3*. The significance of deviations from the symmetrical pattern expected in the absence of gene flow is evaluated using a 'block- jackknife' method to compute standard errrors in genome-wide data (Patterson et al 2012).

The pattern of excess of shared polymorphisms, between Neandertal and non-Africans, observed using this statistical framework was first used as primary line of evidence for the presence of traces of Neandertal admixture in presend-day non-African genomes (Green et al., 2010). However, as initially noted by Green et al. 2010, the test relies on the assumption that the population ancestral to the sampled populations was randomly mating, and later work showed that the pattern of excess of shared polymorphisms between Neandertal and non-Africans is also compatible with a scenario of ancestral subdivision, without invoking admixture (Eriksson and Manica, 2012).

Green and colleagues (2010) also introduced a related statistic to estimate *f*, the proportion of the genome shared through admixture (Green et al., 2010; Durand et al., 2011). This test makes use of the numerator of the *D statistic*, which is named *S* (Equation 1). The value of *S* is divided by the amount of introgression seen in a scenario of complete admixture (i.e. if the population *P2* was interely of *P3* ancestry). The Equation 2 is adapted from Equation S18.4 in Green 2010 in which *P3a* and *P3b* were required to be two different Neandertals.

$$f_{GREEN} = \frac{E[S(P1,P2,P3,O)]}{E[S(P1,P3a,P3b,O)]} \qquad (2)$$

D statistics are very similar to a set of statistics that recently became quite popular as tests for admixture. These tests rely on the *f-statistics* originally proposed by Reich et al (2009), and are based on a statistical framework derived from the study of patterns of allele frequency correlation across populations, designed to measure the proportion of genetic drift shared between populations (Patterson et al., 2012; Peter, 2015). This group of statistical tests includes the *f2*, *f3* and *f4 tests* (Equations 3, 4 and 5, lower case letters refer to allele frequencies in the different populations).

$$f2(P1, P2) = E(p1 - p2)^2 \qquad (3)$$

$$f3(Px; P1, P2) = E(px - p1)(px - p2) \qquad (4)$$

$$f4(P1, P2; P3, P4) = E(p1 - p2)(p3 - p4) \qquad (5)$$

These statistics can easily be interpreted under a population phylogeny, as by definition they make use of the additivity of branch lengths in the tree.

Consider the population phylogeny shown in Figure 2A; *P1, P2, P3* and *P4* are populations labels, with *PX* being a putative admixed population. *f2* can intuitively be interpreted as the paths between two populations (vertices) in the tree; *f3* and *f4* can be considered as external and internal branches of the phylogeny, respectively (Peter, 2015). Then, the expected values of the *f-statistics* can be computed from the overlap of drift paths in the phylogenetic tree, while in the presence of admixture the drift follow different paths (see Patterson et al., 2012, Appendix A).

Here we give a more detailed illustration of the *f4-ratio estimation* and of one of its recent application for estimating Denisovan ancestry (Reich et al., 2011). Further information about *f2* and *f3* derivation can be found in Patterson et al. (2012).

Also referred as "f4 ancestry estimation", the *f4-ratio estimation* was initially defined in Reich et al. (2009) to provide evidence of admixture in Indian populations, and since then has gained extensive popularity. The ratio (Equation 6) measures the ancestry proportions in an admixed population, under the assumption that the correct demographic model for the population contributing to the statistic is known (Patterson et al., 2012).

Consider the phylogeny shown in Figure 2B. Let *PX* be the population for which we are estimating the admixture proportion, an estimate of α can be obtained as:

$$\alpha = \frac{f4(P1,P4;Px,P3)}{f4(P1,P4;P2,P3)} \qquad (6)$$

where *P2* and *P3* are the populations we assume to contribute to *PX* with proportions α and *1-* α, respectively; *P1* and *P4* are populations with no contribution to *PX* (Figure 2B). In particular, *P1* represents a population more closely related to *P2*, while *P4* is an outgroup (Peter, 2015).

A *f4-ratio statistic* in the form shown in Equation 7 is an example of application of this statistic to genome-wide SNP data to investigate which modern human population in Southeast Asia and Oceania have inherited genetic material from Denisovans (Reich et al., 2011).

$$\alpha_D = \frac{f4(Outgroup, Denisova; East\ Asian, Px)}{f4(Outgroup, Denisova; East\ Asian, New\ Guinea)} \qquad (7)$$

Specifically, the computed statistic quantifies the proportion of Denisova ancestry in each population under study, as a fraction of that found in New Guineans.

"Fossil free" methods for archaic admixture inference.

The study of patterns of linkage disequilibrium (LD) in modern human genomes, represents an alternative method for identifying archaic introgression, using only sequence polymorphism data from extant humans. The method relies on the calculation of *S*,* a summary statistic that looks for population-specific variants that are in strong LD with each other. This approach, introduced by Wall (2000) and Plagnol and Wall (2006), exploits the fact that admixture predictably results in the presence of long divergent haplotypes at low frequency.

*S** is designed to detect highly divergent haplotypes harbouring supposedly 'archaic' variants that are in strong LD and are not found in a "reference" population, i.e. a population not expected to contain any level of archaic introgressed sequences. The method has the advantage of not relying on limited and error-prone ancient DNA sequences, and has the potential to discover and characterize (unknown) archaic human groups, that admixed with early modern humans, in absence of ancient DNA sequence from these archaic species.

On the other hand, when putative archaic introgressed haplotypes in a specific population are found to have high levels of similarity with an outgroup population or species, this signal might not necessarily be caused by introgression from the outgroup. An alternative explanation is ancestral shared polymorphism due to ancestral population structure. This model claims that a haplotype existed before the divergence from the outgroup and only survived in a particular population and the outgroup, but not in other populations (Racimo et al., 2015). However, in the case of admixture with Neandertals, if such admixture events occurred more recently than the divergence with modern humans, we would expect archaic introgressed haplotypes to be shorter. Therefore, a way to distinguish ancestral shared haplotypes from introgressed haplotypes is by measuring their length (Racimo et al., 2015). DNA from Neandertal introgression events should fall into longer contiguous tracts than DNA resulting from ancestral population structure (Mendez et al., 2012).

Using *S** on sequence data from 222 genes, Wall and Hammer (2009) could detect low levels of introgression in East Asians from an unknown hominin source. The authors speculated that this signal might come from admixture with hominin species thought to have co-existed with modern humans in Southeast Asia, such as *H. erectus* and *H. floresiensis* (Swisher et al. 1996; Brown et al., 2004; Morwood et al. 2004).

With a similar statistical framework, highly divergence haplotypes have also been found in 61 non-coding autosomal regions in a sample of three sub-Saharan African populations, that included Mandenka, Biaka, and San, suggesting that archaic forms of the genus *Homo* still undescribed, probably admixed with anatomically modern humans within

Africa (Hammer et al., 2011). When whole-genome sequences from Neanderthals and Denisova became available, using the S^* statistic in combination with the Patterson's *D statistic*, the same authors were able to show that Neanderthals contributed more DNA to modern East Asians than to modern Europeans (Wall et al., 2013).

However, most of the above mentioned studies had the limitation of relying on an implementation of S^* that could highlight evidence of introgression but did not identify the specific putative introgressed haplotypes or which individuals carry them.

With the advent of high-quality sequences, a genome-wide S^* scan implementation applied to 15 African hunter-gatherer genomes (Western Pygmy, Hadza, and Sandawe) led to the discovery of low levels of introgression from a yet undetermined archaic human group, providing evidence for an even more complex history of hominin interactions in Africa (Lachance et al., 2012).

More recently, Vernot and Akey (2014) extending previous work (Lachance et al. 2012), used S^* in a two-stage computational framework to search for surviving Neandertal lineages into non-African whole genomes data from the 1000 Genomes Project. A schematic overview of the newly developed framework is shown in Figure 2C.

Briefly, the first stage of the approach consists in the identification of candidate introgressed sequences using a windowed S^* score calculation. S^* is calculated on individuals and then a hypothesis test is performed to determine whether any individual have a S^* large enough to reject the null hypothesis that they do not carry introgressed sequence (see below). The statistic for the i^{th} individual in a region is calculated as $S_i^* = \max_{J \subseteq V_i} S(J)$, where:

$$S(J) = \sum_{j \in J} \begin{cases} -\infty, & d(j, j+1) > 5 \\ -10000, & d(j, j+1) \in \{1 \dots 5\} \\ 5000 + bp(j, j+1), & d(j, j+1) = 0 \\ 0, & j = max(J) \end{cases}$$

V_i is the set of all variants in individual i in this region, and J is a subset of those variants. Variants that are also found in the reference population are not included in this analysis. In the calculation of *S*, we treat *J* as a list of variants ordered by genomic position. Thus, the variants j and $j+1$ denote adjacent variants in *J*. *d(j,j+1)* is the genotype distance between two variants, where genotypes are coded as 0, 1, and 2, and the distance between two variants is the sum of the difference between their genotype values in each individual. The term *bp(j,j+1)* is the distance in base pairs between two variants. In the calculation of *S(J)*: -∞

does not allow consecutive variants in *J* to have more than five genotype differences; -10000 is a penalty for consecutive variants with 1-5 genotype differences; variants with no genotype differences (perfect LD) are scored 5000 + the distance between them, which gives a higher score to variants in perfect LD that extend over larger distances; the final line allows the last variant to be added.

The stage described above allows for the discovery of a subset of variants in high LD within a window. The process of calculating S* makes use of a dynamic programming method computing the S score for all possible pairs of variants, with the goal of finding the subset (*J*) of variants in that individual (V_i) that maximizes *S(J).*

The subset of variants that maximizes *S is* then considered for inclusion in the putative introgressed haplotype. Notably, the first stage (highlithed in red in Figure 2C) does not rely on the use of ancient DNA sequences from archaic species.

In a second stage, the set of candidate variants with top *S** score is then refined through a direct comparison to the Neandertal genome, i.e. testing whether the variants identified in the first stage match Neandertal significantly more than expected by chance (Vernot and Akey, 2014). This study showed that fragments of Neandertal ancestry can be robustly identified in the genome of modern human non-Africans populations to an extent that a substantial amount - corresponding to ~20% of the Neandertal genome - can be resurrected from the genome of living humans (Vernot and Akey, 2014).

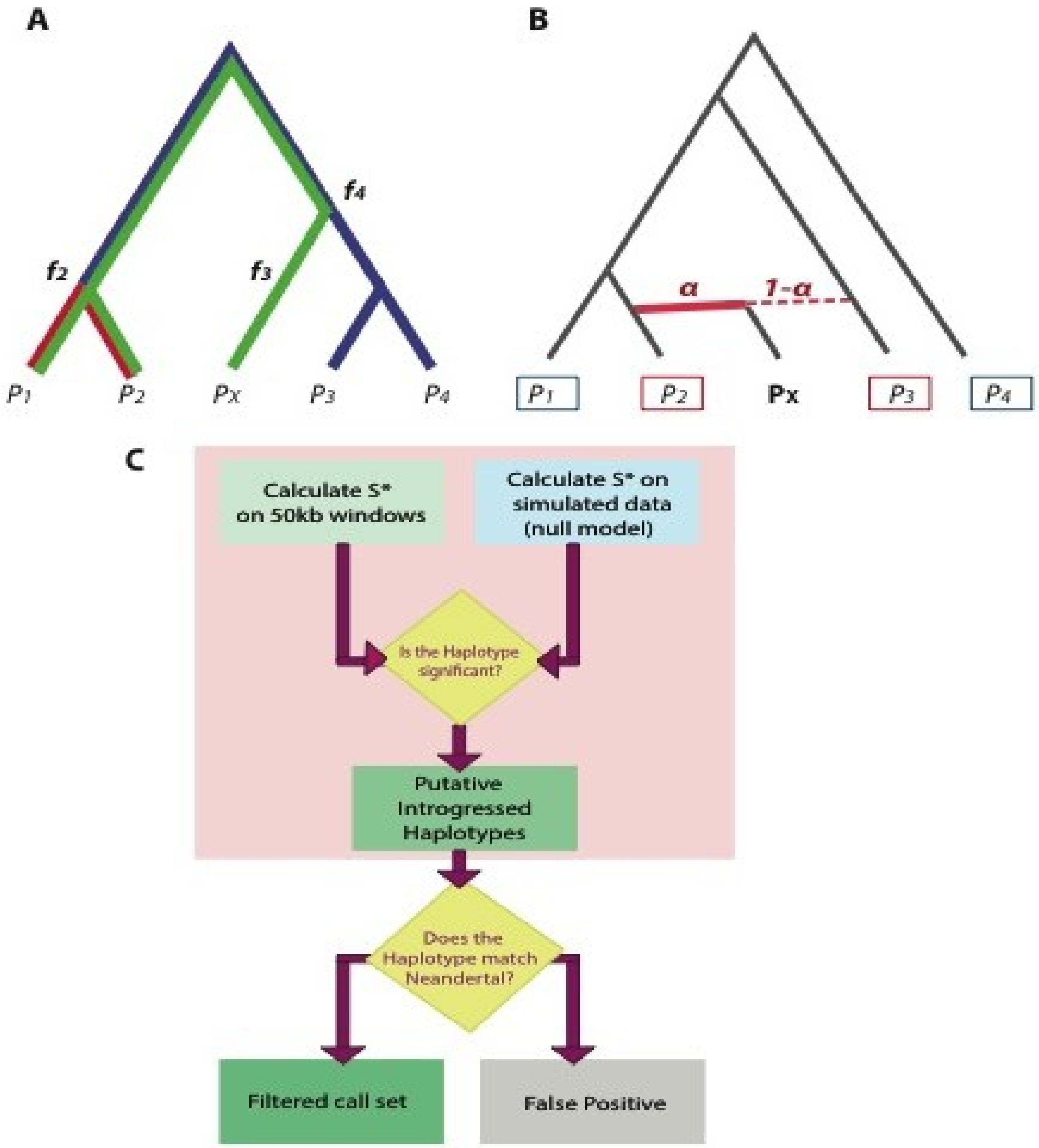
A
f4
f2
f3
P1
P2
PX
P3
P4
B
α
1-α
P1
P2
Px
P3
P4
C
Calculate S* on 50kb windows
Calculate S* on simulated data (null model)
Is the Haplotype significant?
Putative Introgressed Haplotypes
Does the Haplotype match Neandertal?
Filtered call set
False Positive

Figure 2. A). Schematic visualization of a population phylogeny to show how f-statistics can be seen as a measure of shared genetic drift between two, three and four populations. Different colors refer to the branch lengths tested using the f2, f3 and f4 tests.
B) Model underlying the f4-ratio estimation (modified from Peter 2015). C) S* statistical f ramework as developed by Vernot and Akey (2014). Modified from Vernot and Akey 2014.

Archaic legacy

The advent of high-quality genome data from archaic species not only offers new insights into their interactions with modern humans, but also allows us to start to investigate in detail the content of these introgressed regions and the possible consequences for the human populations that inherited them.

Recently there has been increasing evidence that supports the presence of genetic variants, acquired from archaic humans, that could be beneficial to modern humans. This process is known as *adaptive introgression*, and is likely to have played an important role in the interaction with new environments during the human exodus from Africa, such as the exposure to new pathogens or to different climatic conditions.

Several cases of adaptive introgression have been described so far. For example, a recent study showed that an haplotype introgressed from Denisova in *EPAS1*, a gene that encodes a transcription factor involved in the response to hypoxia, might be responsible for conferring altitude adaptation to Tibetans (Huerta Sanchez et al., 2014). Another study by Vernot and Akey (2014) found that the gene *BNC2* could be a candidate for adaptive introgression from Neandertal in Europeans. The introgressed haplotype, presents at high frequencies in Europeans but absent in Asians, harbors a SNP associated with skin pigmentation in Europeans that might have helped modern humans to face lower sunlight exposures, potentially leading to vitamin D deficits, in non-African environments. Interestingly, this introgressed region has been also described in another independent study (Sankararaman et al., 2014).

While there is evidence that admixture with Neandertlas might have been selectively advantageous to the modern human gene pool - on the other hand, the presence of large genomic regions strongly depleted of Neandertal ancestry, seems consistent with the action of recurrent selection against deleterious sequences that entered the modern human gene pool thorugh admixture with archaic hominins (Vernot and Akey, 2014). Interestingly, a large depletion on 7q of Neandertal lineages has been found in East Asians and Europeans (Vernot and Akey, 2014), encompassing the *FOXP2* locus, a tran-scription factor that plays an important role in human speech and language which has been associated with speech and language (Enard et al., 2002).

Research aims

The aim of this project was to collect genetic samples from the pygmy population living in the Rampasasa village in the island of Flores (Eastern Indonesia). Using newly available high-throughput genome sequencing technologies and "fossil free" methods for archaic admixture inference recently developed by Joshua Akey and his lab at the University of Washington (Seattle), we aimed to use these genomic data to infer the demographic history of the natives of Rampasasa. The main goal was to detect putative archaic introgressed regions in their genomes, if any. Our Null hypothesis is that the genomes of the short statured pygmy inhabitants of Rampasasa show no evidence of hybridization with archaic human forms. The alternative hypothesis is that they are, in fact, descendants of the hybridization between *H. floresiensis*, or a related archaic hominin form, and modern humans who later migrated into Flores.

METHODS

Samples collection

The Flores samples considered in this study were collected from the village of Rampasasa in the Manggarai District (Flores, Eastern Indonesia). Approval for this study was obtained from the Human Subjects Review Committee at UCSC to Principal Investigator Richard Edward Green (UCSC Institutional Review Board), following local approvals from Dr. Herawati Sudoyo, Deputy Director of Eijkman Institute for Molecular Biology of Jakarta and from the Manggarai District authority in Ruteng (Flores).

A courtesy visit was made to the village in Fall 2013, during which we gave full explanation of the project aims and sample collection procedure. Upon approval of the elders' committee of Rampasasa and on the community level, samples were collected from 32 individuals in collaboration with the Eijkman Institute in Spring 2014. Informed consent, written in their own languages and in English, was obtained from all participants

Saliva samples were collected using the Oragene DISCOVER (OGR-500) DNA sample collection kits (Genotek, Ottawa, Ontario, Canada). Individuals were selected trying to maximize the number of unrelated individuals with the shortest stature, designated as ''Pygmies'' (or its literal translation in local languages). Anthropometric measures (stature), self-reported ancestry information, clan and language were also collected in the field. DNA was extracted using Qiagen High Molecular Weight Blood and Tissue kit for DNA extraction. See Supplementary Table S1 for information about samples collected.

Selecting individuals for whole-genome sequencing

Genome-Wide SNP Genotyping

All samples were genotyped on the Illumina HumanOmni2.5-8 v1.1 BeadChip genotyping array. Genotypes were called using the Illumina Genome Studio v2011.1, genotyping Module Version 1.9.4 for a total number of 2,391,739 SNPs, which includes

2,336,044 autosomal SNPs, 53,260 X-chromosome, 2,246 Y-chromosome and 189 mitochondrial SNPs.

PLINK version 1.9 (Purcell et al., 2007) was used to assess genotyping quality according to the protocol published by Anderson and colleagues (2011). Samples were checked for outlying heterozygosity (more than 3 standard deviations from the mean) and elevated rates of missing data (genotyping failure rate >3%). Genotype data from the X-chromosome was used to check for discondance with ascertained sex and highlight potential plating errors.

Identification of duplicated samples or cryptic relatedness

To infer relationships among individuals, we employed the software KING version 1.4 (Manichaikul et al., 2010). We checked family relationship by estimating the kinship coefficient (the probability that two alleles sampled at random from two individuals are identical by descent) using KING robust algorithm that allows the existence of population structure (through parameter *--kinship*) on 2,336,044 autosomal SNPs. The analysis was performed on the unpruned dataset, following KING documentation. Pairwise relationships were checked between each pair of individuals. Pairs of individuals with estimated kinship coefficients between 0.177 - 0.354, 0.0884 - 0.177 and 0.0442 - 0.0884 were considered first-degree, second-degree, and third-degree relationships, respectively (Manichaikul et al., 2010).

The proportion of the SNPs at which there were 0, 1, and 2 shared alleles identical-by-decent (IBD) — denoted by Z_0, Z_1, and Z_2 respectively—was also analyzed using the method of moments (MoM) implemented in the software package PLINK on each pair of 32 individuals. After filtering (--geno 0.1 and maf 0.05), 1,125,329 SNPs autosomal SNPs were used in the analysis. The IBD analysis was used to assist validating the pairs of first-degree of relationships inferred by KING. Without genotyping errors and mutations, Z_0 and Z_1 of monozygotic twins or duplicate sample pairs are expected to be 0 and Z_2 to be 1; Z_0 and Z_2 of PO pairs are expected to be 0 and Z_1 to be 1; Z_0 and Z_2 of FS pairs are expected to be 0.25 and Z_1 to be 0.5. Therefore, the proportion of IBD (denoted by PI_HAT = P (IBD = 2)+0.5*P (IBD = 1)), which equals to or above 0.5, suggests that the pairs are close relatives. Those pairs with PH_HAT value greater than 0.44 (0.05 genotyping error rate and 0.01 mutation rate were assumed) denote close relatives.

A *NxN* matrix of genome-wide IBS pairwise distances, was also estimated using the R package SNPRelate (Zheng et al., 2012).

Finally, to better visualize the results of the relationship inference obtained using KING and PLINK, a *NxN* matrix of the pairwise relationships between the 32 individuals was represented in a network, using the R package "igraph".

Deconvoluting population affinities in Asia

To focus on population affinities within Asia, we integrated our Flores genotypes with SNP data released by the HUGO Pan-Asian SNP Consortium (Abdulla et al., 2009), genotyped for Affymetrix Genechip Human Mapping 50K Xba array and available at http://www4a.biotec.or.th/PASNP. After removing duplicates and close relatives within the Pan-Asian dataset as reported in (Yang et al., 2011), we merged the dataset with our Flores genotypes. The presence of duplicated individuals between our Flores sample and the Pan-Asian dataset was checked using KING. PLINK was used for data managment and quality control. Genotyping success rate was set to 95% and MAF to 0.01. The full merged dataset included 17,035 SNPs genotyped in 1,685 individuals from 74 populations (Supplementary Table S2). After LD-pruning (--indep-pairwise 50 5 0.4), a Principal Component Analysis (PCA) was calculated on 15,187 SNPs using the R package SNPRelate.

Analysis of whole-genome sequencing data

Whole-Genome Sequencing and Filtering

We generated high-coverage sequences from approximately 1µp of genomic DNA for the 10 Flores selected samples. All sequencing work has been carried out at the New York Genome Center (NYGC). Sequencing libraries were prepared using TruSeqDNA Nano 350bp kits and 150 bp paired-end reads were obtained on a HiSeq X Ten sequencing platform. Sequenced reads were aligned to the human reference sequence GRCh37 (available from GATK bundle ftp://ftp.broadinstitute.org/bundle/) using the BWA-MEM algorithm (Li and Durbin, 2009, http://bio-bwa.sourceforge.net). Duplicates reads were removed using Picard (http://broadinstitute.github.io/picard/). Local realignment around indels and base quality score recalibration were performed using GATK (McKenna et al., 2010) to generate the final bam files.

Coverage depth was calculated using *bedtools genomecov* (Quinlan et al., 2010). We used reads mapped to the Y- chromosome to highlight potential plating errors.

Variants were called following GATK3.2-2 pipeline (see GATK “Best Practices” documentation and DePristo et al 2011 for more details). Calls were obtained on each single sample using GATK HaplotypeCaller (default filters were used: Minimum mapping quality = 20, Minimum depth of coverage = 10, Maximum depth = 500, Minimum base quality = 10) and per-sample gVCFs were generated. Subsequently, joint genotyping was performed using GATK GenotypeGVCFs. Variant calls were annotated with SnpEff (Cingolani et al., 2012) and v*cftools* (Danecek et al., 2011). rsIDs were retrieved from dbSNP Build 141 available at

ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b141_GRCh37p13/VCF/All.vcf.gz

All analysis were restricted to SNPs, and unless stated, were carried out using the following filters. First, we identified reliably callable loci from aligned reads with GATKCallableLoci considering the following thresholds:

- Minimum base quality of 20
- Minumum mapping quality of 30
- Minimum depth of 10
- Maximum depth equal to the 99.5th percentile of autosomal depth, calculated for each sample

To ensure overlap among sites for all the individuals, we only considered sites that passed filters in all the individuals.

Then, the following sites were masked:

- sites within 5 bp of a short insertion or deletion;
- sites where every individual was heterozygous as in Lachance et al. (2012);
- sites with FisherStrand (FS) more than 60, as recommended by GATK Best Practice;
- sites within segmental duplications (Bailey et al., 2002; downloaded from: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/genomicSuperDups.txt.gz
- sites where sites where at least 18 of 35 overlapping 35-mers from the human reference sequence can be mapped elsewhere with zero or one mismatch. (Li and Durbin, 2011)
- sites within a CpG dinucleotide context as in Prufer et al. (2013);
- sites included in the 1000 Genomes accessibility mask, downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/20141020.pilot_mask.whole_genome.bed. This mask was applied in the analyses perfomed in the context of the 1000 Genomes dataset.
- sites included in the Altai and Denisovan minimal filter mask (Prufer et al., 2014), downloaded from: https://bioinf.eva.mpg.de/altai_minimal_filters/

To further validate the high-quality of our calls, genotypes obtained from the sequencing data were then compared with those from the HumanOmni2.5-8 v1.1 BeadChip genotyping array, using GATKGenotypeConcordance. After Ignoring A/G and C/T sites to avoid strand flipping, and multiallelic sites, calls for a total of 1,069,991 were compared.

Inference of haplotype phase

Genotypes were computationally phased using the algorithm implemented in Beagle version 4.0 (Browning and Browning, 2007) and using a panel of 2,504 computationally phased genomes from the 1000 Genomes Project, as reference. Beagle forms a HMM by locally clustering the haplotypes at each marker position along a chromosome and uses stochastic expectation maximization (EM) to converge towards the most probable solutions (Browning and Browning, 2011). The phasing was performed per chromosome for all the autosomes in two steps. Namely, (1) we first phased sites in the Flores individuals that were also present in the 1000 Genomes dataset, using 1000 Genomes sites as reference panel; (2) sites not present in the 1000 Genomes dataset were then phased using the phased Flores sites as reference panel. Finally, we merged those sets of phased sites. A trio was included in this process as there is a significant number of haplotypes shared among father-mother-child, and this information facilitates phasing of unrelated individuals from the same population. We then merged the fully phased Flores VCFs with the 1000 Genomes phased VCFs. Any sites that were a) unmasked in either the Flores or 1000 Genomes datasets, and b) present in one dataset but absent in the other, were assumed to be homozygous reference.

Integration of WGS with Human Origin dataset

We obtained a published population dataset genotyped on the Affymetrix Human Origins SNP (for a technical description of the array see Patterson et al. 2012 and Lazaridis et al. 2014). Briefly, the array was designed as a union of 13 different SNP panels for which the ascertainment is well understood (Keinan et al., 2007; Patterson et al., 2012) . Twelve panels include heterozygous sites discovered within 12 individuals of known ancestry, sequenced in two recent papers (Green et al., 2010; Reich et al., 2010), while the 13th panel includes sites where a randomly chosen allele is derived in a San individual relative to both Denisova and chimpanzee. The 12 modern human samples used in the discovery panels are all from the CEPH-HGDP panel.

The fully public curated dataset was downloaded from http://genetics.med.harvard.edu/reich/Reich_Lab/Datasets_files/EuropeFullyPublic.tar.gz. The files in PACKEDANCESTRYMAP formats, were converted to PLINK formats using the "CONVERTF" utility of ADMIXTOOLS software version 3.0 (Patterson et al., 2012).

The Human Origin dataset is released in a publicly curated version from which all individuals and all markers loci failing QC were previously removed, as described in Lazaridis et al. (2014). The public dataset also includes sequencing data from 5 archaic human samples (Vindija Neandertal, high coverage Altai Neandertal, low and high coverage Denisovan and the Mezmaiskaya Neandertal), 11 ancient humans, the human reference genome (hg19) and 5 Primates samples.

We next merged the Flores sequencing data with the genotyping dataset. We started by extracting positions from our multi-VCF that overlap with loci genotyped in the Affymetrix Human Origin SNP Array, whereas positions not called in the Flores multi-VCF were set as homozygous for the reference allele. Only variant and homozygous sites that passed our callable loci filters were used in the Flores samples (see Whole-Genome Sequencing and Filtering).

The so obtained Flores callable multi-VCF, filtered for high confidence positions genotyped in the Human Origins Dataset, was converted to PLINK formats using v*cftools* (Danecek et al., 2011). The Flores genotypes were then merged with the Human Origin array data. After discarding sex-linked, mitochondrial and multiallelic sites and SNPs with ambiguous strand identification in the Human Origins dataset, 593,269 autosomal SNPs are left for the analysis.

After removing populations with less than three individuals from the published dataset, and related individuals, our final analysis data could be based on 1,925 present day humans from 159 worldwide populations, 5 archaic human samples, and the chimpanzee sequencing data. A list of the samples included in the dataset is provided in Table S4, together with geographic coordinates that were used for the map in Figure 6A.

This merged filtered dataset (which we shall refer to as “HO Dataset”), was used for subsequent analyses including Principal Component Analysis (PCA), ADMIXTURE, inbreeding analysis and estimates of ancestry proportions using *D* and *f4-ratio statistics*.

Inference of population structure

To explore the Flores genome diversity in the context of worldwide variation, we performed a Principal Component Analysis (PCA) with the merged HO Dataset (above). The original merged dataset was pruned using PLINK in order to avoid the effect of variants in high linkage disequilibrium, employing a window of 200 SNPs advanced by 25 SNPs and a r^2 threshold of 0.4 (*--indep-pairwise* 200 25 0.4), matching parameters previously used (Lazaridis et al., 2014). After removing SNPs with a missing genotype rate >= 0.05 and SNPs with MAF <= 0.05, the PCA was carried out using the R package SNPRelate on 163,746 SNPs.

We used the unsupervised clustering algorithm ADMIXTURE (Alexander et al., 2009) to infer ancestral clusters in our Flores samples and 158 worldwide populations of the Human Origin dataset. In order to avoid the effect of variants in high linkage disequilibrium, we pruned our dataset using PLINK following the same approach as above.

We ran ADMIXTURE in 10 replicates with different random seeds, with a 5-fold cross-validation error, exploring number of clusters (K) ranging from 2 to 7. The multiple runs were then aligned using the "greedy “algorithm of CLUMPP (Jakobsson and Rosenberg,

2007) and visualized with the software *Distruct* (Rosenberg, 2004). Finally, we computed a *f4-ratio statistic* in the following form to estimate the amount of "Papuan" ancestry in our Flores individuals.

$$\alpha_{PAPUAN} = \frac{f4(Yoruba, Papuan; Han, Flores)}{f4(Yoruba, Papuan; Han, Australian)}$$

Estimates of Inbreeding

Runs of homozygosity were detected using PLINK applying a sliding moving window of 5000 kb (minimum 50 SNPs) across the genome. We allowed one heterozygous and five missing calls per window. To exclude runs that are likely to be affected by strong LD, a threshold of 500kb was set for the minimum length needed for a tract to qualify as homozygous, since in most human populations LD tipically extends for relatively short distances (Abecasis et al., 2001; Reich et al., 2001).

PSMC analysis

We used the Pairwise Sequentially Markovian Coalescent (PSMC) to infer long-term effective population sizes in our Flores individuals (Li and Durbin, 2011). The PSMC was applied to 9 unrelated Flores pygmies, along with 11 previously published high coverage genomes (Meyer et al., 2012, Prufer et al., 2014). Alignments to the hg19/GRCh37 were downloaded from http://cdna.eva.mpg.de/denisova/. Diploid consensus sequences were generated for each of the 20 individual using the 'pileup' command of SAMtools (Version: 1.2) (Li et al., 2009). The following sites were marked as missing as recommended in Li and Durbin (2011):

- sites where read depth is higher than 60 or below 10. Such thresholds, that approximate 2 times and 1/3 of the average depth respectively, were chosen to account for the lower average depth in the published HGDP genomes (~30x) compared to the Flores genomes;
- sites where the root-mean-square mapping quality is below 10;
- sites within 5bp of a short insertion or deletion;
- sites where the estimated consensus quality is below 30;

- sites where less than 18 out 35 overlapping 35-bp oligonucleotides from the human reference sequence, can be mapped elsewhere with zero or one mismatch.

The PSMC analysis was performed using default parameters. Results were scaled assuming a mutation rate of 1.25 x 10^{-8} per site per generation (Scally and Durbin, 2012) and assuming 25 years per generation.

Inference of Denisovan and Neandertal ancestry in the Flores pygmies.

To explore the relationship between modern humans and archaic hominins, we first performed a PCA on the Altai Neandertal, Denisovan, and chimpanzee genome, included in the merged HO Dataset using the R package SNPRelate (Zheng et al.,2012), and projected 1,925 present-day humans onto the plane described by the top two principal components.

We then applied formal tests of admixture, such as the *D-statistics* and the *f4-ratio* statistics (Reich et al., 2009; Green et al., 2010; Patterson et al., 2012) to verify the presence of Neandertal and Denisova ancestry in the Flores genomes. We converted PLINK formats to eigenstrat formats using the "CONVERTF" utility of ADMIXTOOLS v.3 (Patterson et al., 2012). To detect gene flow between modern humans and Neandertals, for each population in the HO Dataset, we calculated a *D-statistics* in the form *D(X, Yoruba, Altai Neandertal, Chimp*) using ADMIXTOOLS, which computes a standard error using a weighted block jackknife for each estimated quantity (block size set to 5 cM). In order to contrast the genetic similarity of modern humans with the two archaic species, we calculated a *D statistics* in the form of *D(Yoruba, X, Altai Neandertal, Denisova).*

Finally, the proportion of Denisova admixture in the Flores genomes was estimated using the *f4-ratio* statistics in the form *f4(Yoruba, Altai Neandertal; Han Chinese, Flores)/f4(Yoruba, Altai Neandertal; Han Chinese, Denisova),* using the “qpF4ratio” software of ADMIXTOOLS, which computes a standard error using a weighted block jackknife for each estimated quantity (block size set to 5 cM) (Patterson et al., 2012).

Statistical method to identify and classify archaic sequences

In order to identify archaic sequences in the genomes of the Flores pygmies, we extended the framework previously described in Vernot and Akey (2014). Specifically, we used a three-stage approach to *(1)* identify candidate introgressed sequences using the S^* statistic (Plagnol and Wall 2006; Vernot and Akey 2014), *(2)* calculate a *p-value* to quantify whether a putatively introgressed haplotype matched Denisovan or Neandertal sequence more than expected by chance, and finally *(3)* refine and probabilistically classify the set of haplotypes.

Notably, the method adopted in Vernot and Akey (2014) was developed to identify sequences inherited from Neandertal ancestors in modern human populations. Here we used an extended version of the framework, designed to detect both Neandertal and Denisovan sequences in the Flores individuals (Vernot, Tucci et al., 2016). Thus, our framework makes inferences from the bivariate distribution of archaic match *p-values.* The inference steps are described in detail below. Please notice that here we aim to identify putative introgressed sequences that were inherited from an archaic species, *H. floresiensis*, for which there is no genome available. To this purpose we first proceed with the identification of sequences that were instead inherited from Neandertal and Denisova. This allows us to separate the archaic signals deriving from Neandertal and Denisova admixture, from the signal that might come from admixture with *H.floresiensis* or other unknown sources.

Step 1: Calculating S^*

We used the S^* framework, as described in Vernot, Tucci and colleagues (2016), for initially identifying putatively introgressed archaic sequences. This step does not use *any* information about available archaic reference genomes (Vernot and Akey, 2014; Lachance et al., 2012), and thus is applicable to identifying introgressed sequence from an unsequenced hominin such as *H. floresiensis*. S^* is designed to detect divergent haplotypes whose variants are in strong linkage disequilibrium and are not found in a "reference" population. 107 Yoruban genomes were used as a reference population in the calculation, as we do not expect to find variants introgressed from *H. floresiensis* in Africa.

S^* operates on a single Flores individual, combined with the reference panel of Yoruban individuals. The following simplifies the S^* for the ith individual in a sample to $S_i^* = \max_{J \subseteq V_i} S(J)$:

$$S(J) = \sum_{j \in J} \begin{cases} -10000, & d(j, j+1) > 0 \\ 5000 + bp(j, j+1), & d(j, j+1) = 0 \\ 0, & j = max(J) \end{cases}$$

Where V_i is the set of all variants in individual i in this region, and *J* is a subset of those variants. Variants that are also found in the reference population are not included in this analysis. In the calculation of *S(J)*, we treat *J* as a list of variants ordered by genomic position. Thus, variants j and $j+1$ denote adjacent variants. The term *d(j,j+1)* represents the genotype distance between two variants, where genotypes are coded as 0, 1, and 2, and the distance between two variants is the sum of the difference between their genotype values in individual i. The term *bp(j,j+1)* is the distance in base pairs between two variants. In the calculation of *S(J)*: -10000 is a penalty for consecutive variants with 1-5 genotype differences; variants with no genotype differences (perfect LD) are scored 5000 + the distance between them, which gives a higher score to variants in perfect LD that extend over larger distances; the final line allows the last variant to be added.

To calculate $S*$, we looked for the set of variants *J* that maximizes *S(J)*, using an efficient dynamic programming algorithm that allows computation of $S*$ in genome-wide datasets. See also Vernot, Tucci and colleagues (2016) for futher details.

We then estimated a null distribution of $S*$ values by simulating sequence data using *ms* (Hudson, 2002), under demographic model from Vernot and Akey (2014) and Vernot, Tucci et al. (2016). We simulated under a grid of recombination rates and population diversity, and build a generalized linear model to the grid of $S*$ quantiles using the R package *mgcv* (Wood, 2011*),* as described in *(*Vernot and Akey, 2014). For each putative introgressed haplotype, we then used this model to estimate the $S*$ percentile based on the population diversity and recombination rate, and retained putative introgressed haplotypes with an $S*$ score in the 99th percentile of null simulations, obtaining our $S*$ callset.

Step 2: Calculating archaic match p-values

We now considered the $S*$ callset for each Flores individual, which is statistically enriched for archaic sequences but has not been compared to any archaic genome, and calculate archaic match *p-values* against both Neandertal and Denisovan in a method similar to that described in (Vernot and Akey, 2014). We generated an empirical distribution of the expected archaic match percentage (the number of matches to the given archaic / the number of variable sites) in an African population, without substantial Neandertal introgression, given the characteristics of the putative introgressed haplotype, and used this distribution to obtain an empirical archaic match *p-value* as in Vernot, Tucci et al (2016).

Step 3: Calling and classifying archaic sequence

We used a likelihood method as developed in Vernot, Tucci et al., (2016), which operates on the bivariate distribution of Neandertal and Denisovan match *p-values,* and also leverages simulations with and without archaic admixture. Specifically, the method estimates for the set of *S** significant haplotypes, the proportion of Neandertal, Denisovan and "*unknown*" sequences, i.e., not introgressed from Neandertal or Denisova but harboring *S** significant haplotypes - and converts archaic match *p-values* into posterior probabilities, allowing us to identify introgressed haplotypes at a desired FDR and probabilistically assign them the labels of "Neandertal", "Denisovan", or "*unknown*". Please notice that the *unknown* category is expected to harbor archaic sequences with unknown origin (not Neandertal or Denisova), as well as false positives. For this analysis we used a threshold for Neandertal and Denisovan match *p-values* such that FDR = 5% (i.e., 5% of these calls are expected to be non-Neandertal and non-Denisovan – potentially from other archaic sources, and potentially non-introgressed). *Unknown* haplotypes were further refined selecting all the haplotypes which showed a Neandertal match p-value >.1 and a Denisova match p-value >.1. We estimate a FDR for this *unknown* callset of < 7% (i.e., < 7% of these haplotypes derive from Neandertals or Denisovans).

We ran the same S* statistical framework on 501 European, 502 East Asian and 482 South Asian individuals from the 1000 Genomes Project, and 35 high coverage Melanesian genomes from Vernot, Tucci et al. (2016), and compared the amounts of *unknown* sequences found in these populations and in the Flores genomes.

Estimating ages of unknown introgressed sequences

We used an algorithm for "ancestral recombination graph" (ARG) inference, implemented in the program ARGweaver (Rasmussen et al., 2014). An ARG provides a record of all coalescence and recombination events since the divergence of the sequences under study, and specifies a complete genealogy at each genomic position.

We applied ARGweaver to two datasets: the "original" dataset consisting of the Altai Neanderthal and Denisovan, six present-day humans from Africa and chimpanzee (panTro4), and the "original+Flores" consisting of our Flores genomes in addition to the previous mentioned genomes. The six present-day humans included are two Yorubans (HGDP00927, SS6004475), two Mbuti (SS6004471, HGDP0456), and two San (HGDP01029, SS6004473).

ARGweaver was run on the set of *unknown S** significant regions, for which we required Neandertal and Denisovan match p-values to be > .1. Using posterior probabilities calculated as in Vernot, Tucci et al., (2016), we estimated that this subset contains 93% true non-Neandertal or Denisovan haplotypes – corresponding to an FDR of 7%. These *unknown* regions were padded by 100kb and merged. We then ran ARGweaver on each region for 2,000 MCMC iterations, with the first 500 iterations discarded as burnin. In addition to the filters described in Kuhlwilm et al., (2016), we filtered our Flores data for callable loci, regions around indels, sites with excess of heterozysity, FisherStrand, to minimize influence from alignment and genotype-calling errors (see Materials and Method for details about our filtering strategy).

Sampled ARGs were summarized in terms of the time to most recent common ancestor (TMRCA) for both the "original" and the "original+Flores" datasets. We proceed selecting haplotypes where the estimated TMRCA for the original dataset was at least 20,000 generations (500ky), and TMRCA was increased by at least 5,000 generations when Flores individuals are added. As we expected to find some levels of variability in TMRCA estimates, especially for short genomic intervals, we attempted to identify longer stretches where the Flores individuals harbor divergent haplotypes by padding all such ancient coalescents by 2kb and then merging these regions.

Finally, we annotated the genes that overlap these regions, using the RefSeq database (Pruitt et al., 2014) and refFlat table obtained from UCSC Genome Browser (Karolchik et al., 2004).

RESULTS

Selecting individuals for whole-genome sequencing

Genotype quality and cryptic relatedness

To describe genomic variation in Flores and investigate the presence of genomic regions that might be inherithed from *H. floresiensis*, we sampled 32 pygmy individuals from Flores and produced whole genome sequences from this population (see Table S1 for sample information).

Given the limited information about the origin of the individuals sampled and the high costs associated with the sequencing of the entire set of collected samples, we set up a strategy for selecting a subset of individuals for whole-genome sequencing, based on a screening of genotype data to primarily avoid (a) the presence of groups of related individuals; (b) the typing of individuals with uncertain ancestry; (c) the inclusion in the subset of extreme genetic outliers. We then performed SNP genotyping for all our Flores individuals for a total of 2,391,739 SNPs.

First, we evaluated the quality of our data, looking for individuals with outlying heterozygosity and missing genotype rate. The observed heterozygosity rate per individual was calculated and plotted versus the proportion of missing SNPs per individual (Figure 3). Thresholds (dashed lines) were set to ≥ 0.03 for the genotype failure rate and ± 3 standard deviations from the mean for the heterozygosity rate (Anderson et al., 2011). No sample failed these QCs. Samples were also checked for discordant sex information (mismatches between documented sex and that suggested in the genotyping data) to highlight potential plating errors. No individual with discondance sex was identified.

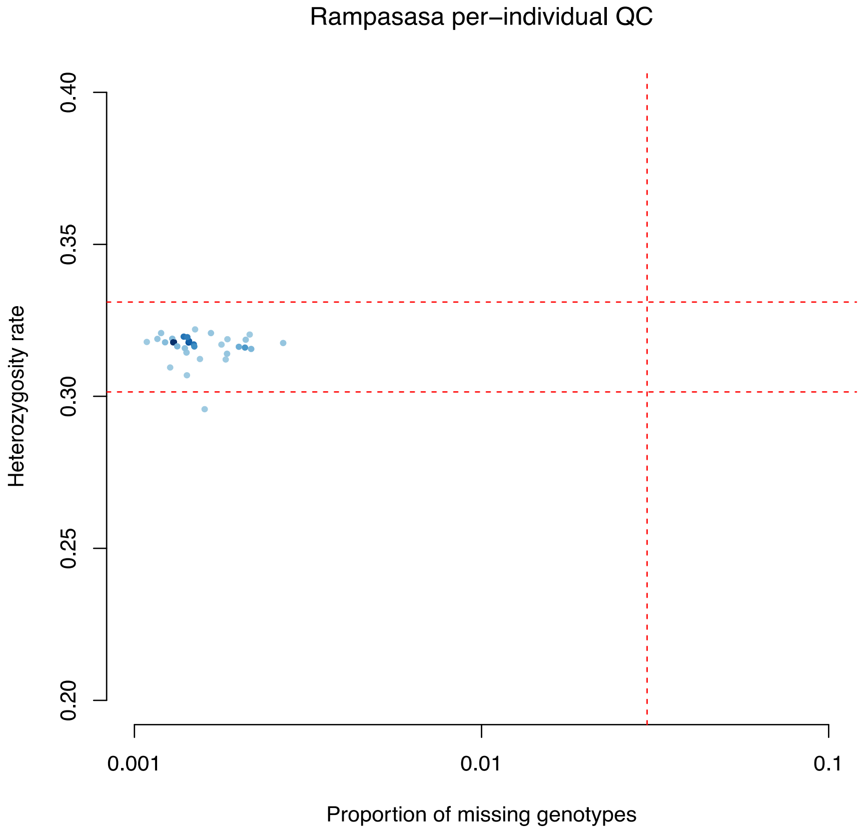

Figure 3. Genotype failure rate vs. heterozygosity across all 32 individuals. Shading indicates sample density and dashed lines denote QC thresholds

Second, given the small size of the village and its relative geographic isolation, we sought to identify close relatives or duplicates among the individuals sampled, so as to later exclude them from the sequencing subset. We then checked family relationships by estimating the kinship coefficient, i.e. the probability that two alleles sampled at random from two individuals are identical by descent, for all pairwise relationships. No duplicated samples were found among the 32 individuals; however we identified 34 pairs of related individuals up to third-degree, which include: 7 pairs of parent-offspring (PO), 2 pairs of full-sibling (FS), 6 pairs of second degree and 19 pairs of third-degree relationships, as inferred by KING (Figure 4A). Dashed red lines correspond to the thresholds for first, second and third degree relationships, respectively. A negative kinship coefficient estimation indicates that two individuals are unrelated and may also point to population structure leading to high levels of diversification between the two individuals. First degree related individuals inferred by KING were also confirmed using the method of moments (MoM) implemented in PLINK. Figure 4B shows the proportion of identity by descent (IBD), denoted by PI_HAT, against the probability of sharing 0 alleles IBD. Pairs of first degree relatives are expected to show values of PI_HAT greater than 0.44 (Purcell et al., 2007). To better detect similarities between pairs of individuals, we also computed a square, symmetric distance matrix based on identity by state (IBS) where the distance between each pair of individuals is reported as the proportion of alleles which are *not* IBS (Figure 4C). The distance between each pair ranges in theory from 0 to 1. In real life, we do not expect to find observed values close to 0, since individuals with no known relationship would anyway share a very large proportion of the genome identical by state because of shared evolutionary history (Barbujani & Colonna, 2010). By contrast, we take values close to 1 (yellow shades) to indicate duplicated samples or first degree relatives.

Taken together these results allowed us to disantagle the presence of cryptic relatedness, i.e. the undocumented presence of close relatives within a sample of ostensibly unrelated individuals.

To better visualize the family relationships within our sample, we drew a network of the pairwise relationships between the related individuals identified in our analysis (Figure 4D). First and second-degrees relatives were then excluded from downstream analysis. In deciding which individual to remove for each pair, we retained individuals with the shorter stature (stature measures are shown in Supplementary Table S1 and in Figure 4D).

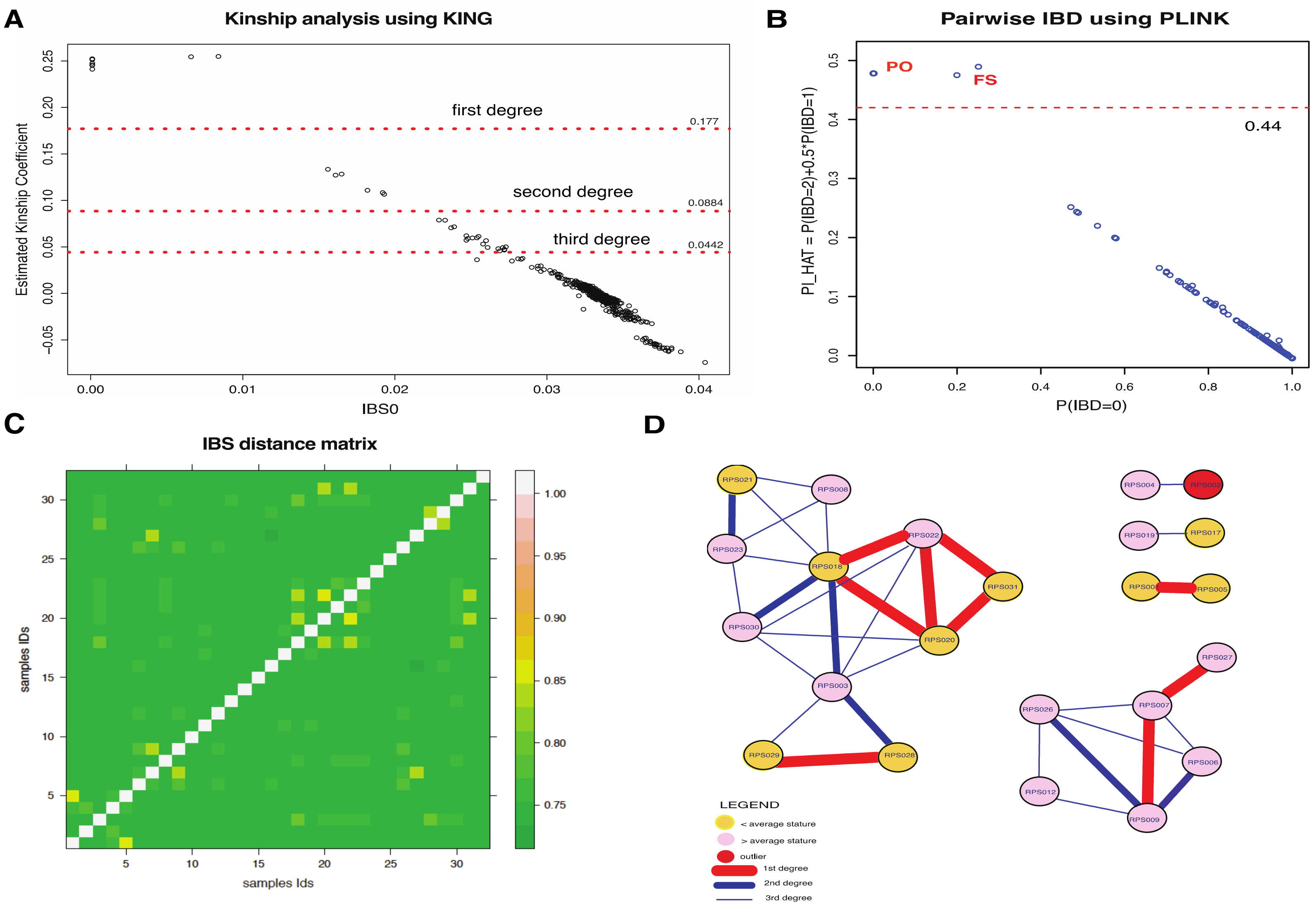

Figure 4. Relatedness analysis. A) Kinship coefficient estimated using KING is plotted against the proportion of zero IBS-sharing. Dashed lines indicate inference criteria reported in KING documentation; B) The proportion of identity by descent (IBD), denoted by PI_HAT, was estimated using the method of moments (MoM) implemented in PLINK and plotted against the probability of sharing 0 alleles IBD; C) Distance matrix based on identity by state (IBS). D) Network of family relationships as inferred in our analysis. Vertices represent individuals, colors are associated to the stature of the individual (red = outlier, yellow ≤ mean stature, pink ≥ mean stature). Edge colors and thickness represent relathionships inferred using KING (see legend). Individual stature is shown in Supplementary Table X. Mean value for males stature (146.52 cm) was calculated excluding the outlier individual which measured 175 cm.

Flores in the context of Asian genetic diversity

At the time of our preliminary analysis for selecting individuals for whole genome sequencing, the best resource of SNP data from the region of our interest was the PanAsia HUGO dataset (Abdulla et al., 2009).

Although limited to ~50,000 markers, the Pan-Asian SNP dataset represents the largest survey of genetic variation in East and Southeast Asia. Among others, the dataset includes 15 populations from the Indonesian Archipelago, for a total of ~300 individuals sampled. Interestingly, 17 samples were collected years ago from the same pygmy population of Flores. Our KING analysis confirmed that 6 of those individuals from the Pan-Asian overlapped with our study, and these were thus excluded from the analysis.

To obtain a first, synthetic view of the main patterns of population affinity in Asia, we performed a PCA (Figure 5A; populations codes refer to Supplementary Table S2). As shown in Figure 5A, the first component clearly distinguished East Asian populations from those of the Indian subcontinent. Along the second component, we observed a north-south gradient of diversity spanning from East Asia to Oceania. All Flores individuals under study (black dots) fell within the range of genetic diversity seen in Indonesian populations, and as previously reported (Abdulla et al., 2009), appeared to be genetically close to populations from Southeast Asia, such as the Malaysian Negritos (light greens) and the Philippines Negritos (dark yellows).

To further refine our analysis, we restricted our dataset to include 1,503 individuals from 64 East and Southeast Asian populations. The PCA in Figure 5B, computed on the reduced dataset, offer a clearer view of population structure in the region. The first component tends to separate continental East Asia from Island Southeast Asia, with individuals from Oceania on the most divergent edge. Indonesian populations (reds, within polygon) are widely spread along the first and second components. Within ISEA, the second component revelaed the presence of two main clusters: one is represented by Malaysian Negritos and Western Indonesians, while the second includes populations from Eastern Indonesia and Philipines Negritos. Within the second cluster, East Indonesian populations tend to form an East-West genetic cline which has been shown to be linked to increasing amounts of “Asian” ancestry going west (Xu et al., 2012). As expected, our Flores samples clustered with the Rampasasa villagers from the Pan-Asia dataset (IDRA, grey dots), along with other populations from Eastern Indonesia. Interestingly, as observed previously by Xu et al. (2012), Flores appeared to be closer to a population from the island of Sumba, than to a neighbour population from Flores island (IDSO).

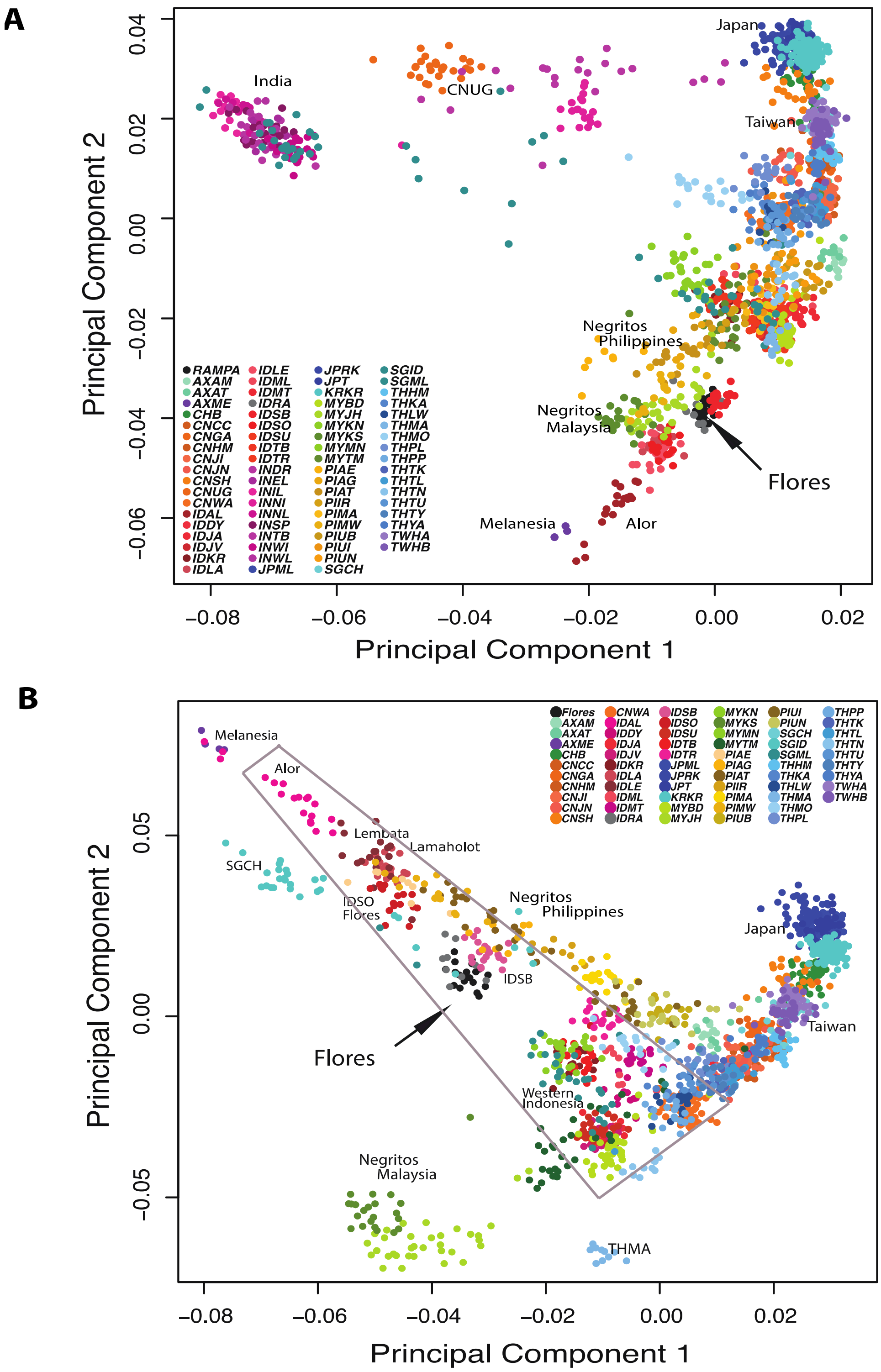

Figure 5. Principal Component Analysis calculated on (A) 74 populations and (B) 64 populations from the Pan-Asian HUGO Dataset. Indonesian populations are shown inside the polygon.

The Flores genomes

After we estimated family relantionships and verified the ancestry of the individuals sampled, we selected 10 individuals for whole-genome sequencing. Our approach tried to maximize the number of unrelated individuals up to third degree, with shortest stature and belonging to Tuke, Taga and Kina, the most common clans among the shortest individuals. We included a trio to facilitate haplotype inference but we considered only unrelated individuals in downstream analyses.

Each of the genomes was sequenced to a median autosomal depth of 37.8x (range 33 – 49x; Table S3 and Figure S1) by Illumina technology. QC-passed mapped reads per individual ranged from 657,389,731 to 100,9711,349 per individual.

We identified 9,871,310 raw variants that included 8,206,481 SNPs and 1,664,829 INDELs. After applying our callable filters and masking complex regions, 5,335,281 autosomal SNPs were left for the analysis (which correspond to ~ 68% of the 7,799,013 autosomal SNP identified). We compared those SNPs with dbSNP Build 141 and found that 252,209 represented new variants discovered in this sequencing project. Of those novel variants, 206,272 were singletons.

When looking at the number of total reads that mapped to the Y chromosome, we noticed incosistencies with the individuals' documented sex (Figure S2). We further investigated this finding using KING, looking at all possible pairwise comparisons within the 10 individuals sequenced and the corresponding genotype data from HumanOmni2.5. We found that a mislabeling error likely occurred during the sequencing so that 2 pairs of individual labels were exchanged. Fortunately, we were able to link the sequencing data to the correct individuals thanks to the genotype data we previously produced.

After fixing the samples labels, we then compared our genotypes obtained from the sequencing data and those from the HumanOmni2.5-8 v1.1 BeadChip genotyping array, and found that the correlation was never below 99.8%.

Flores in the context of worldwide genetic diversity

To date, these data represent the first high coverage genomes from Indonesia and the first genome-scale survey of a pygmy population from Flores.

Although numerous genetic studies have been conducted on pygmy populations, such as the Baka and Bakola in Africa (Batini et al., 2011; Jarvis et al., 2012; Lachance et al., 2012) or the Negritos in Southeast Asia (Jinam et al., 2013; Aghakhanian et al., 2015), nothing is known about the genetic history of the Indonesian pygmies of Flores.

To explore the Flores genomes diversity in the context of worldwide genetic variation, we performed a PCA. We merged our data with the Affymetrix Human Origins SNP dataset (Patterson et al., 2012), an excellent resource that allowed us to make comparisons with genotype data obtained from a worldwide population panel. Figure 6A shows the geographic distribution of the populations analyzed in this study. A list of the samples included in the HO Dataset is provided in Table S4, together with geographic coordinates that were used for the map in Figure 6A.

The PCA was carried out on 1,925 individuals on a set of 163,746 LD-pruned SNPs. As shown in Figure 6B, the first component, which captures 5.86% of total variation, separates Africa from Eurasia and America (for color codes see Figure 6A and Figure S3). Populations in the Middle East, Arabian Peninsula and East Africa are intermediate between the African cluster and the Eurasian cluster. The second component, which accounts for 4.52% of the variation, separates West Eurasia and South Asia from East Asia and the Americas. Our Flores individuals fall close to the cluster of East Asian populations, but showed a tendency toward Oceanian populations.

Model-based clustering analysis was carried out using the maximum-likelihood approach implemented in ADMIXTURE on the LD-pruned HO Dataset. The results of the analysis from K=2 to K=7 are shown in Figure 6C. Each individual is represented by a vertical bar and different colors refer to their estimated ancestry components. Black lines separate distinct populations. To provide a better visualization of the admixture proportions, the Flores samples are shown on the right edge of the plot. At K=2, the inferred ancestry components distinguish African (red) from non-African populations (green). With K=3 we distinguish Asian and American populations (yellow), from West Eurasian. The "Asian" component appears to be dominant in the Flores genomes, as well as in populations from Oceania. Starting from K= 5, Flores along with populations in Oceania (Papuans from New Guinea, Australians and Bouganville), can be distinguished from Asian populations. While this "Oceanian" ancestry component is dominant in Oceania, it appers to be smaller in Flores. For values of K > 5, Flores is characterized by the presence of a dominant Asian-related ancestry component and a small proportion of ancestry shared with populations in Oceania.These results are highly consistent with our *f4-ratio* estimates of an average 35% of "Papuan" ancestry in the Flores genomes (range 31-38%; Table S5), and with previous observations (Xu et al., 2012; Lipson et al., 2014).

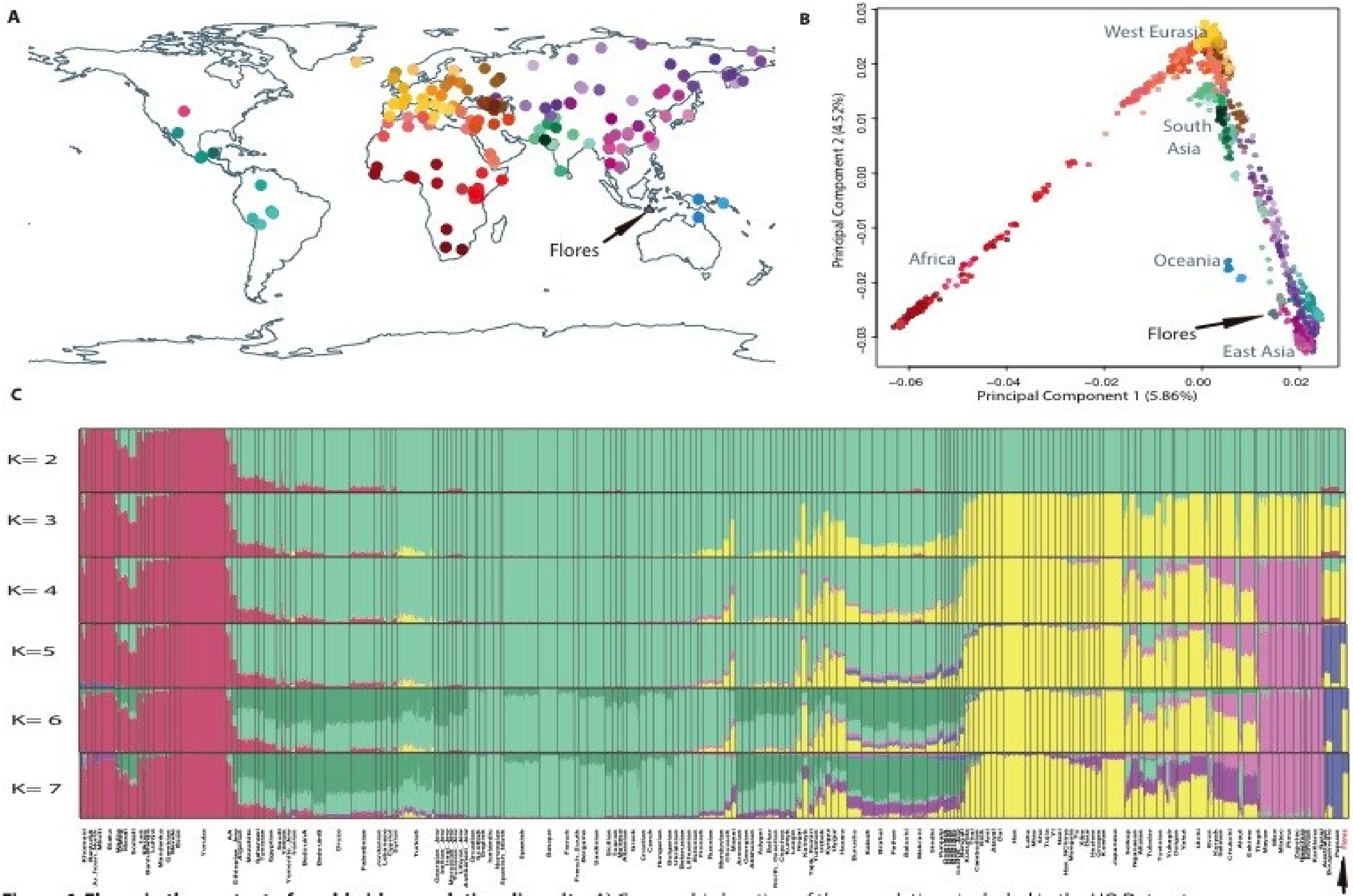

Figure 6. Flores in the context of worldwide population diversity. A) Geographic location of the populations included in the HO Dataset. B) Principal Component Analysis performed on 159 populations included in the HO Dataset. C) ADMIXTURE analysis. Flores individuals are shown on the right edge of the plot.

Inbreeding in the Flores pygmies

Inbreeding occurs when mates are more closely related than expected if chosen at random in a population. “Inbreeding Avoidance” is a taboo present in most cultures (Brown, 1991), and likely evolved as a cultural adaptation to prevent deleterious effects of inbreeding (Pusey and Wolf, 1996). However, in small populations, inbreeding can be the simple effect of random mating with limited possibilities to choose a mate.

Although marriages between close relatives are prohibited in the region, because the Rampasasa village has traditionally been composed of just a few families living in the impenetrable tropical forest in the inner part of the Flores island, it is more than likely that some level of inbreeding might be present.

The inbreeding coefficient of an individual, *F* is defined as the probability that two alleles chosen at random at a homologous locus within an individual, are *identical by descent (IBD)* with respect to a base (reference) population in which all alleles are independent (Wright, 1922). Homozygosity caused by two IBD alleles is termed autozygosity, as opposed to allozygosity, which is homozygosity produced by alleles that are *identical by state (IBS)*. *F* can potentially be estimated from pedigrees, and from genome-wide data, using a marker-by-marker estimate, and runs of homozygosity.

Here we used runs of homozygosity (ROH), i.e. long tracts of consecutive homozygous sites, observable in high-density genome-scan data from the apparently outbred populations included in the HO Dataset, to get a reliable and accurate estimate of autozygosity at the individual level.

Figure 7A shows the distribution of the number of ROHs and their cumulative size (in base pairs) for each individuals in the HO Dataset. Longest tracts are expected in individuals with an appreciable degree of inbreeding. The Flores pygmies (grey triangles) showed ROH lengths comparable to other Asian populations. We note one outlying individual (RPS013) with higher amounts of ROH, which might reflect some level of consanguinity. Exceptionally long and numerous ROH are seen in one individual from the Kusunda people, a Nepalese linguistic isolate, while outlying individuals are also seen among the Native Americans Surui, Mixe, Piapoco and Karitiana, as previously observed (Rodriguez-Flores et al., 2016). African populations have the fewest long tracts per individual, consistent with a larger number of generations, and hence amount of recombination, since their founding (Gibson et al., 2006). Within Africa, we observed individuals with high runs of homozygosity belonging to the Hazda from Tanzania. Previous studies suggested that both population bottlenecks and inbreeding might have contributed to the observed pattern in the Hazda (Henn et al., 2011; Lachance et al., 2012).

We then restricted our analysis to populations from East Asia and Oceania (Figure 7B), and found that our Flores samples (grey triangles), with the exception of individual RPS013, showed ROHs always below 100Mbp. The Flores ROHs are shorter than those from populations in Oceania (blue shades), but longer than most of the East Asian samples (pink shades). However, we do not see in the Flores genomes extreme lengths and amounts of ROHs as we see in the Atayal or in the Ami from Taiwan. Overall ROH lengths in Flores are comparable to those seen in mixed European populations (McQuillan et al., 2008), such as the Orcadians and Dalmatians, and are much shorter compared to those observed in Negritos populatons from Malaysia (Aghakanian et al., 2015).

Inference of long-term effective population sizes

We chose a model-flexible method based on the pairwise sequentially Markovian coalescent (PSMC) framework, to estimate population size changes over time, using the Flores genomes, five high coverage African genomes (YRI, Mbuti, MDK, San, Dinka), and six high coverage non-African genomes (Dai, Sardinia, European, Han Chinese, Karitiana and Papuan). The PSMC approach works by estimating the distribution of the time since the most recent common ancestor of the two haploid genomes carried by a single individual. Figure 7C shows the PSMC outputs for the Flores pygmies (purple shades), and for the other 11 genomes analysed for comparison. We recovered the same patterns as previously observed (Li and Durbin, 2011; Meyer et al., 2012; Prufer et al., 2014; Ayub et al., 2015; Rodrigues-Flores et al., 2016). Going backwards from 150 kya, all genomes seem very similar in their estimated *Ne* history. The African genomes diverge from non-African genomes around 100-120 kya, when non-African, included Flores, experienced a severe bottleneck. Estimates of *Ne* for recent times were not considered here, since it has been shown that the PSMC tends to produce biased estimation for recent population histories (Sheehan et al., 2013; Liu and Fu, 2015). In the most recent lapse of time, fluctuations in the estimated population size in Flores (consistent across all individuals) closely resemble those observed in Papua New Guinea.

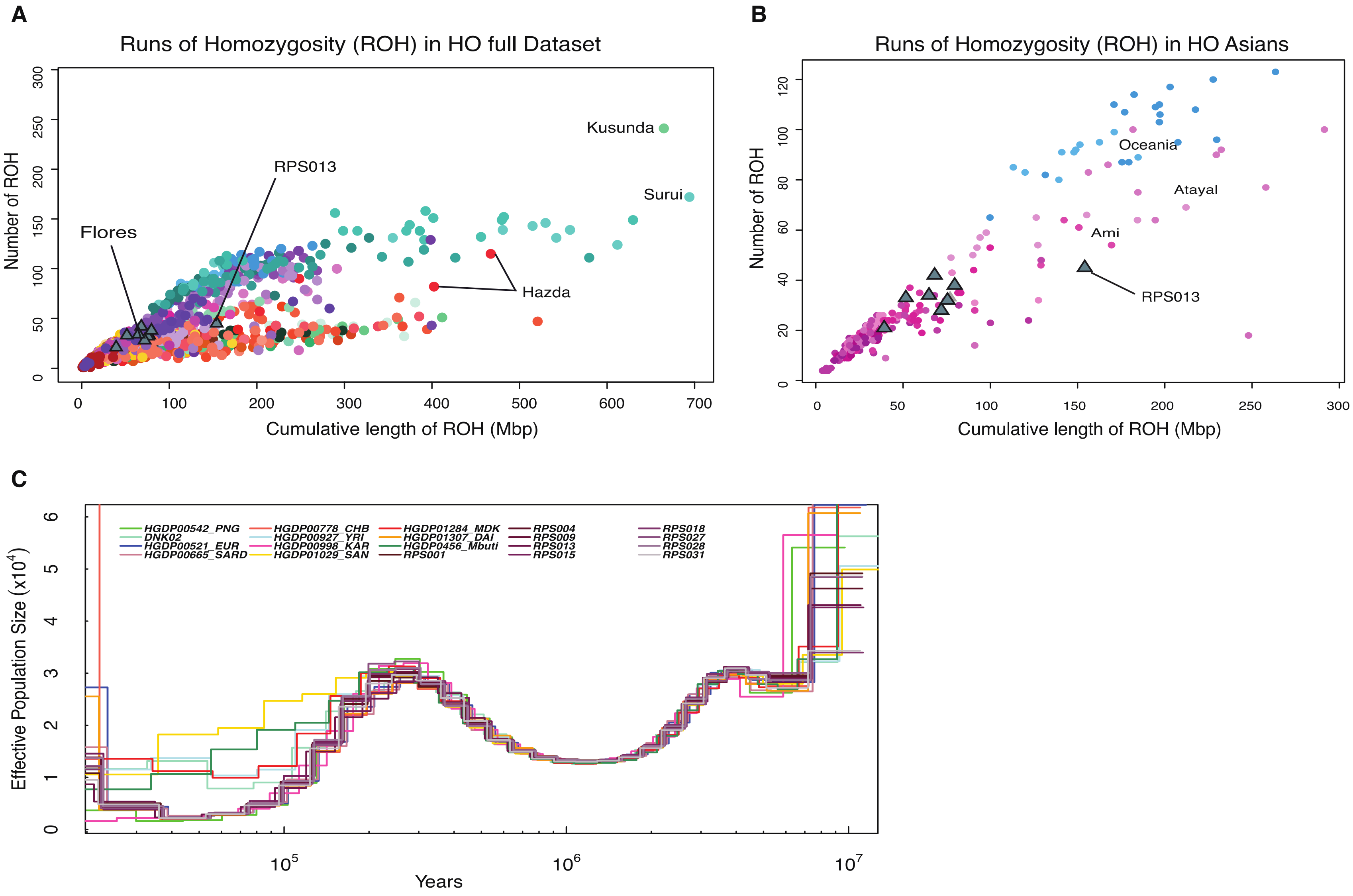

Figure 7. A) Runs of homozygosity A) in the full HO Dataset and B) in populations of Asia and Oceania. Grey triangles correspond to Flores individuals. C) Inferred population size changes over time inferred using the PSMC in 9 unrelated Flores pygmies and 11 published high coverage genomes from worldwide populations. Estimates were scaled assuming a mutation rate of 1.25 x 10−8 per site per generation (Scally and Durbin, 2012) and 25 years per generation.

Neandertal and Denisovan gene flow in the Flores pygmies.

To describe the relationships between modern humans populations and archaic species, we used a PCA projection analysis where we defined the first two components using the Altai Neandertal, the Denisovan, and the chimpanzee genomes and projected 1,925 present-day humans into the resulting axes of variation (Figure 8A). This method, described in detail in Patterson et al. (2006), enables to circumvent the effect of geography on modern human variation, and should in stead capture heterogeneities in their similarity to the archaic humans genomes (Reich et al., 2010, Skoglund and Jakobsson, 2011).

To allow a better visualization, we enlarged the central portion of the PCA showed in Figure 8A, and plotted the mean values for the top two principal components (PCs), for each of the 159 populations (Figure 8B). Under the assumption of no admixture between modern humans and archaic species, we would expect modern humans to be homogeously distribuited in the plane described by the top PCs of archaic hominins and chimpanzee (Skoglund and Jakobsson, 2011). The first component describes the genetic similarity of modern humans to both archaic species, while the second component contrasts modern humans with respect to their similarity to Neanderthals and Denisovans. As we can see in Figure 8B, Eurasians forms a cluster that appears to be more similar to Neandertal, while populations from Oceania tend to be closer to Denisova, recapitulating previously reported observations (Reich et al., 2010, Skoglund and Jakobsova, 2011, Qin and Stoneking, 2015). East Asians and Native Americans appear to be the closest populations to Neandertal, as previously reported (Skoglund and Jakobsson, 2011, Qin and Stoneking, 2015). Interestingly, Flores, is located in close proximity of the Eurasian cluster but unlike other Eurasian populations, it shows a tendency toward Denisova. These results suggest that, in addition to ancestry from Neandertals, Flores might harbor some amount of Denisova ancestry.

To confirm our inference of genetic similarities between modern humans and archaic species, we applied formal tests of admixture. We then tested for the presence of gene flow from Neandertals in our Flores population, using a *D statistics* in the form *D(X, Yoruba, Altai Neandertal, Chimp),* where *X* is a target population in the merged HO Dataset, and compared the value of *D* with those obtained for all the 158 populations in our dataset. We did not compute the *D statistics* using other African populations as outgroup, as it has been extensively shown that the use of different African groups does not produce noticeable effects on ancestry estimates in tests for archaic admixture (Reich et al., 2011; Wall et al., 2013; Qin and Stoneking, 2015).

Our estimates of D for all populations (Table S6 and Figure S3) are consistent with previous estimates (Meyer et al., 2012; Prufer et al., 2014; Qin and Stoneking, 2015). Values plotted in Figure 8C are referred to populations from Asia, Oceania and the Americas, included in the merged HO Dataset. All those populations show large and significant D values (Z-score ≥ 5; Reich et al., 2009; Pattersonet al., 2012) with East Asians and Native Americans showing higher amounts of Neandertal ancestry compared to South Asians and West Eurasians, as previously reported (Wall et al., 2013; Prufer et al., 2014, Qin and Stoneking, 2015). Interestingly, Flores show a value of D higher than other populations from East Asia, but close to the value seen in the Miao from China, that have been previously reported to have high amount of Neandertal ancestry (Qin and Stoneking et al., 2015). The highest D values are observed in populations in Oceania. However, this signal of higher D values in Oceania, as well as in Flores, is more likely due to the presence of Denisova ancestry in these populations (Reich et al., 2011, Qin and Stoneking et al., 2015). A high value of D is also observed in the Native Americans Surui (Figure 8C). This results is consistent with a recent study that suggests that Amazonian Native Americans descend partly from a Native American founding population which carried ancestry closely related to indigenous Australians and New Guineans (Skoglund et al., 2015). Therefore, the excess of archaic signal in the Surui might be due to Denisova ancesty from Oceania, than to an excess of Neandertal admixture (Skoglund et al., 2015).

Given the difficulties to distinguish between the putative genomic contributions coming from Neandertal and Denisova, because of the high similarity between these archaic genomes (Prufer et al., 2014), we then contrasted the genetic similarity of modern humans with these two archaic species using a *D statistics* in the form *D(Yoruba, X, Altai Neandertal, Denisova)* (Table S7). Using this implementation, populations closer to Neandertal are expected to show more negative values of D, compared to populations which harbor Denisova ancestry in addition to Neandertal ancestry, as previously observed by Qin and Stoneking (2015). Our results show that Flores' value of D is more positive compared to other Asian and Native American populations. Although the value of D seen in Flores is not as positive as those seen in Oceanian populations, the signal is indicative of the presence of some amounts of Denisova ancestry in the Flores genomes.

Following this observation, we proceed by estimating the amount of Denisova ancestry in the Flores pygmies, using a *f4-ratio* statistics in the following form:

$$\boldsymbol{\alpha_D} = \frac{f4\ (Yoruba, Altai\ Neandertal; Han\ Chinese, Flores)}{f4(Yoruba, Altai\ Neandertal; Han\ Chinese, Denisova)}$$

We used this *f4-ratio* to estimate Denisova admixture proportions, using Han Chinese to correct for levels of Neandertal ancestry in the Flores pygmies, as implemented in Qin and Stoneking (2015). This *f4-ratio* implementation relies on the observation of the presence of similar amounts of Neandertal ancestry in East Asians and populations in Southeast Asia and Oceania (Qin and Stoneking, 2015). Our estimates of Denisova ancestry in the Flores pygmies ranges from 0.7% to 1.5% (Z-score ≥ 2) in six individuals, and are consistent with the results from the PCA archaic projection and the *D statistics* (Figure 8B and 8C), while *D* values with Z-scores ≤ 2 are observed in 3 individuals. As expected, the proportion of Denisova ancestry in Flores is lower than the proportion observed in the Papuans ~3.4% (Reich et al., 2011; Qin and Stoneking, 2015). Previous studies suggested that the amount of apparent Denisova ancestry in Eastern Indonesia, as well as in Oceania, could actually be consequence of admixture with New Guineans (Reich et al., 2011; Qin and Stoneking, 2015), who appear to carry in their genome a higher fraction of DNA of putative Denisovan origin.

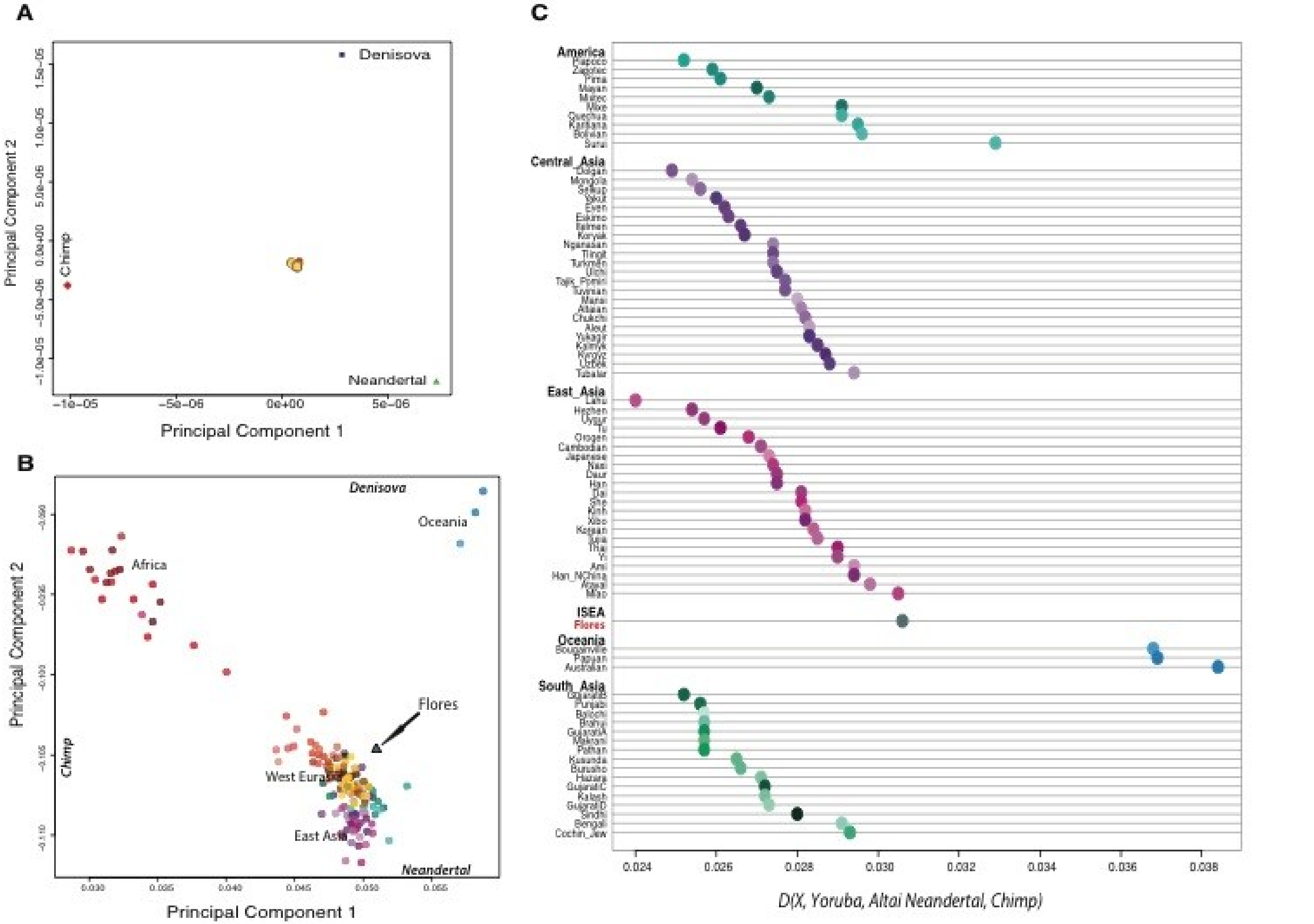

Figure 8. A) Principal Component Analysis to investigate genetic similarities of present-day humans and archaic species. Axes of variation resulting from the PCA performed on the Altai Neandertal, the Denisovan and the chimpanzee, with all 159 modern human pppulations projected into the resulting PCA space. B) Analogous to A) but "zoomed". Mean values per population are plotted. C) D statistics calculated in the for D(Yoruba, X, Neandertal, Denisova), to test for the presence of gene flow from the Altai Neandertal in populations in Asia, Oceania and Americas. Populations are grouped by geographic region, and by ascending values of D.

Ancient introgressed haplotypes in the Flores genomes

The absence of ancient genomic data from *H. floresiensis* limites our ability to make a direct comparison between archaic and modern human genomes using formal tests of admixture, such as *D statistics*. We therefore used the S* framework, as described in Vernot, Tucci et al. (2016), for identifying putatively archaic introgressed sequences, which does not use any information about available archaic reference genomes.

On average, introgressed haplotypes are expected to have an older TMRCA compared to non-introgressed genomic regions and to exhibit high levels of divergence. In the specific case of admixture with *H. floresiensis*, because this hominin species is thought to be descendant of the Asian *H. erectus* lineage - which diverged at least 1 Mya and possibly more from the common ancestor of Neandertal, Denisova and modern humans - we expect to find putative *"floresiensis"* haplotypes to be more divergent compared to modern humans, Neandertal and Denisova haplotypes. Moreover, because admixture with hominins in Flores would have occurred relatively recently – presumably after early modern human dispersal in Island Southeast Asia ~ 40-50 kya - the introgressed haplotype are expected to persist over sizeable genomic regions.

We ran our S* statistical framework on the Flores genomes, 501 European, 502 East Asians and 482 South Asian individuals from the 1000 Genomes Project, and 35 high coverage Melanesian genomes from Vernot, Tucci et al., (2016), and compared the amount of *unknown* sequences found in the Flores genomes with those found in these worldwide populations. Specifically, we recovered on average 225 Mbp of *S** significant haplotypes per individual for our Flores samples, and then identified introgressed haplotypes at a 5% FDR and probabilistically assigned them the labels of "Neandertal", "Denisovan", or "*unknown*", with the last category to include haplotypes that harbor *S** significant haplotypes, but do not match Neandertal or Denisova. *Unknown* haplotypes were further refined selecting all haplotypes with Neandertal match p-value >.1 and a Denisova match p-value >.1, which on average corresponds to ~140 Mb of *unknown* sequences per Flores individual (range 129-151 Mb per individual). Using posterior probabilities calculated as in Vernot, Tucci et al. (2016), we estimated that this refined haplotypes set contains 93% true non-Neandertal or Denisovan haplotypes - meaning a FDR of 7% for the *unknown* category.

Figure 9A shows the amounts of *unknown* sequences versus the amounts of *S** significant haplotypes in the Flores genomes and in the populations used for comparisons. As expected, the two variables are correlated. Melanesians, known to harbor high amounts of Denisova ancestry in addition to Neandertal ancestry, show the highest amounts of *unknown* sequences, likely due to the presence of archaic sequences that did not pass our archaic match thresholds, and might contribuite to inflate the *unknown* category. Interestingly, our Flores individuals seem to have slightly more *unknown* sequences than expected, given the total amount of S* sequence identified, consistent with the hypothesis that these individuals harbor archaic introgression from an unsequenced hominin group.

Because the *unknown* category is likely to harbor some amount of false positives, our main concern was to be able to discriminate between them and the putative true archaic signal in the Flores genomes. We therefore investigated our refined set of *unknown* haplotypes using ARGweaver, a novel Bayesian method to sample full genealogies and corresponding recombination events (ancestral recombination graphs, ARG), which describe the evolutionary relationship between genetic samples (Rasmussen et al., 2014). We applied this method to two datasets: the "original" dataset, consisting of the Altai Neanderthal and Denisovan, six present-day humans and chimpanzee (panTro4) and the "original+Flores" consisting of our Flores genomes in addition to the previous mentioned genomes. An advantage of explicitly sampling ARGs is that it enables inference about local features of the ARG (Rasmussen et al., 2014). Specifically, from the ARG we can extract local trees for every non-recombinant block of the genome, and explore these trees for signs of introgression.

For this analysis, we specifically looked for haplotypes identified in the Flores genomes using our *S** framework, and for which the Flores lineage is inferred to coalesce beyond the African and archaic (Neandertal and Denosova) tree – since these haplotypes are likely to be candidate for introgression from *H. floresiensis (*Figure 9D*)*. Indeed, *true H. floresiensis* introgressed haplotypes are expected to have TMRCA significantly older than the majority of non-introgressed haplotypes, included false positives within the *unknown* category (Figure 9D).

Given the high computational cost of this inferencial method, we initially focused on a subset of the genome (chromosomes 1-9). Our analysis revealed that older TMRCA are obtained when Flores individuals are added to the analysis. In Figure 9B-C we show the distribution of TMRCA for segments of the genomes with length > 100bp, estimated for both the original and original+Flores datasets. Both distributions are characterized by the presence of three main peaks, although one fourth peak, at ~2.3 My, is only present in the Original+Flores distribution. As we can see in Figure 9B, the distribution of TMRCA ages for the original dataset is highly enriched for regions with TMRCA ~700 ky old, with most of the TMRCA falling between 650 and 950 ky, which seems to be consistent with the observation of deep ancestral haplotypes in the Neandertal and Denisova genomes which coalesce above the modern human subtree (Kuhlwilm et al., 2016). The median value of TMRCA for the original dataset is 800 ky (Figure 9C).

When the distribution of the original+Flores is considered, no TMRCA younger than 650 ky is observed. The distribution of TMRCA shows a peak at 1My, two times higher than in the original dataset (Figure 9B). We also observed an enrichment of TMRCAs between 1.4 and 1.5 My, and a high peak which corresponds to a TMRCA of ~1.5 Mya. Interestingly, a fourth small peak unique to the Original+Flores, corresponds to TMRCA estimates of ~2.3 My. The median TMRCA estimated for the Original+Flores is 1 My, older than the one obtained for the original dataset – which seems to be consistent with the hypothesis that the *S** unknown haplotypes might be enriched for haplotypes inherithed through admixture from a divergent group.

In order to identify longer stretches that might harbor an archaic signal, we merged and padded the regions that showed oldest TMRCA, and identified a total of 23 introgressed haplotypes longer than 30 kb. The full list of regions is reported in Table 1, with corresponding gene contents.

Chrom	Chrom Start	Chrom End	Length (bp)	Gene content
chr1	11161920	11267359	105439	LOC105376736,***MTOR***, MTOR-AS1, ANGPTL7
chr1	145613088	145692119	79031	***NBPF10***, RNF115
chr5	99385509	99442000	56491	AC113407
chr3	16595959	16648489	52530	***DAZL***
chr3	17600599	17651899	51300	TBC1D5
chr1	228140720	228191399	50679	CICP26
chr4	150210289	150258719	48430	////
chr6	81486579	81533640	47061	////
chr6	101664359	101710309	45950	////
chr3	127868019	127912139	44120	RUVBL1, EEFSEC
chr8	82519849	82560129	40280	IMPA1P, NIPA2P4, SLC10A5P1
chr1	173984619	174024229	39610	RPS26P34
chr3	170667509	170706269	38760	***MYNN,*** LRRC34
chr3	87797179	87834699	37520	RP11-451B8.1
chr1	112160417	112195969	35552	***RAP1A***
chr1	177332449	177367569	35120	LOC102724661
chr5	92405669	92440720	35051	////
chr1	198555010	198589009	33999	RP11-553K8.2
chr3	106678189	106712000	33811	LINC00882
chr2	224375539	224407589	32050	////
chr3	132821259	132852129	30870	TMEM108
chr6	16988199	17018969	30770	////
chr3	169501099	169531549	30450	***MYNN,*** LRRC34

Table 1. Top longest putative introgressed haplotypes for which the estimated TMRCA for the original dataset was at least 20,000 generations (500ky), and TMRCA was increased by at least 5,000 generations when Flores individuals are added.

Interestingly, we identified a 105 kb long haplotype in chromosome 1 (Figure 9E) overlapping with *MTOR* (mammalian Target of Rapamycin), an evolutionarily conserved serine/threonine kinase that regulates cellular metabolism in response to multiple environmental stimuli, such as growth factors, amino acids, and energy (Laplante et al., 2012). *MTOR* has also been shown to control organismal size, and be responsible for reduction in brain size in other species (Oldham et al., 2000; Long et al., 2002; Cloetta et al., 2013), .

Among those long introgressed regions, we also found a 79kb haplotypes (Figure 9E) which encompasss *NBPF10* (Homo sapiens neuroblastoma breakpoint family, member 10). This gene is a member of the neuroblastoma breakpoint family (*NBPF*) which consists of dozens of recently duplicated genes primarily located in segmental duplications on human chromosome 1. Notably, this gene family has experienced its greatest expansion within the human lineage and has expanded, to a lesser extent, among primates in general. Members of this gene family are characterized by tandemly repeated copies of the *DUF1220* protein domains, located in the chromosomal region 1q21.1, which has been implicated in a number of developmental and neurogenetic diseases such as for example microcephaly, macrocephaly, autism, schizophrenia, mental retardation (Dumas et al., 2012). Given the presence of numerous segmental duplications in the region, we caution that estimates of TMRCA might be affected by the highl complexity and variability of this region.

The gene *DAZL* (Deleted In Azoospermia-Like), encompasses a 52.5 kb haplotype in chromosome 5 (Figure 9E). An autosomal homologue of *DAZ* found in the Y-chromosome, *DAZL* plays a central role during spermatogenesis. Diseases associated with *DAZL* include male sterility and azoospermia (Seboudn et al., 1997). Another gene of similar size (~56 kb) is found in chromosome 5 overlapping with *AC113407*, for which annotation is not available.

We also found a haplotype 35.5 kb long encompassing *RAP1A,* a gene which encodes for a protein of the Ras-related protein family involved in cell growth, development, and plasma membrane-bound signaling.

Finally, the gene *MYNN* (Myoneurin), which belongs to the BTB/POZ and zinc finger protein family implicated in regulatory functions of gene expression, is found in two haplotypes of 38.7 and 30.4 kb respectively. Myoneurin has been identified in various tissues, but muscle is a privileged site of myoneurin gene transcription, where this protein acts as synaptic gene regulator (Alliel et al. 2000).

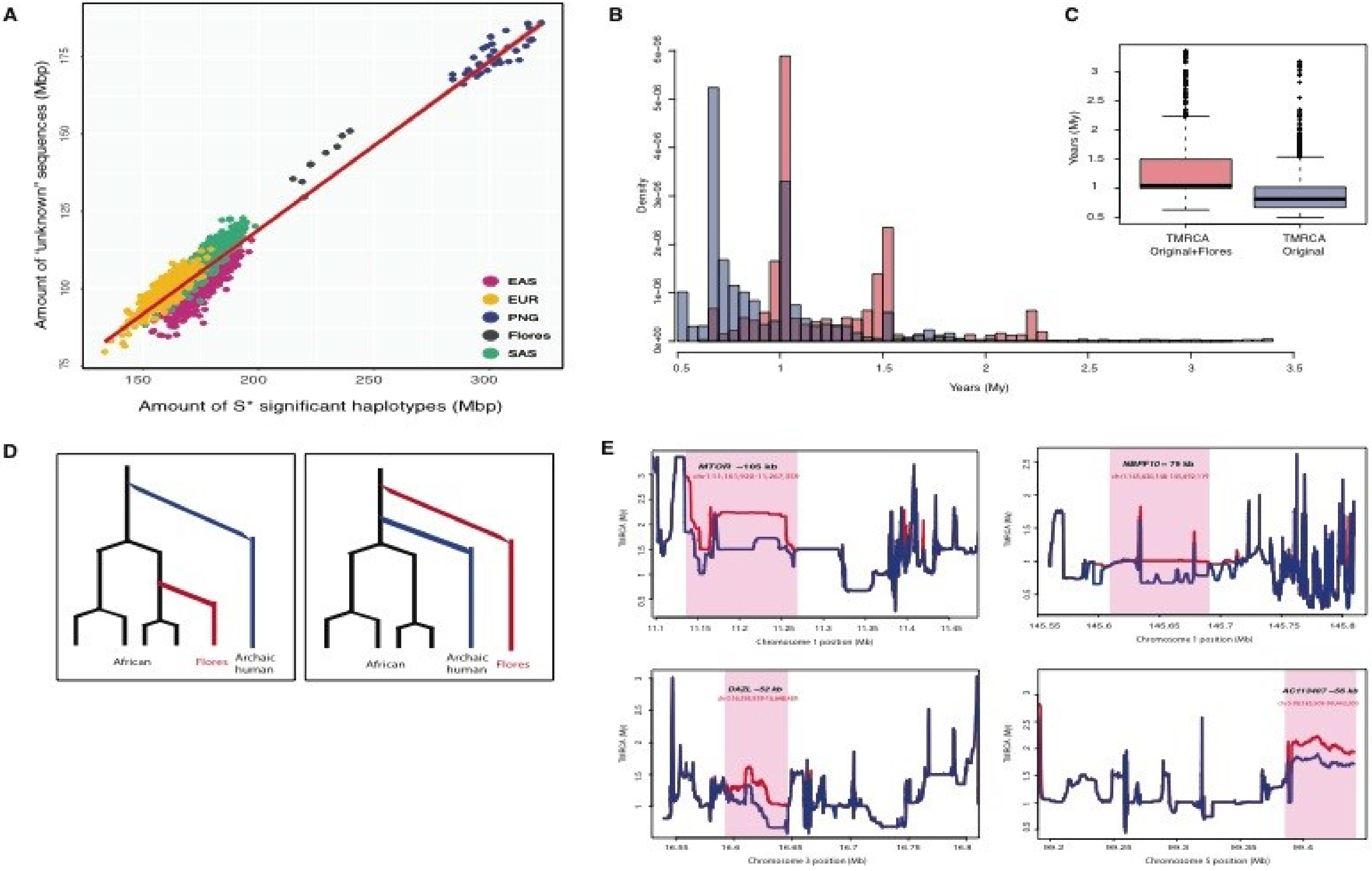

Figure 9. A) Amounts of unknown sequences versus the amounts of S* significant haplotypes in the Flores genomes and in populations used for comparisons (EUR, European; SAS, South Asians; EAS, East Asians; PNG, Melanesians). B-C) Distribution of the TMRCA for haplotypes where the estimated TMRCA for the original dataset was at least 20,000 generations (500ky), and TMRCA was increased by at least 5,000 generations when Flores individuals are added. Time in years is estimated using a generation time of 25 years. D) Distinguishing between two scenarios of non-introgression (left) and introgression (right). E) Posterior mean TMRCAs in the original dataset (blue line) and in the original+Flores (red line) along a genomic regions (Mb). Here we reported the top longest haplotypes merged and padded.

DISCUSSION

Recent revelations that modern human genomes contain traces of introgression though admixture with archaic species, such as Neandertals and Denisovans, suggest that there were not insurmountable barriers to gene flow between hominin groups that lived in the past.

As the landscape of Neandertal and Denisova admixture is being drawn, these new understandings have motivated further exploration for traces of admixture with other extinct hominins, such as the so-called "Flores Man".

H. floresiensis' discovery in the island of Flores in Eastern Indonesia, has taken evolutionary biologists on a very unexpected journey, not only for its astonishing morphology - but also for its surprising age. Indeed this hominin group is thought to have lived as recently as 18,000 years ago, overlapping in time with modern humans in Island Southeast Asia.

Besides the numerous questions that this discovery raised, we aimed to investigate whether *H. floresiensis* genes survive in the genomes of contemporary humans.

Here we present ten high quality genome sequences of a pygmy population recently discovered in the island of Flores. This population lives in a village close to the Liang Bua cave, where the *H. floresiensis* fossils were found and has attracted the interest of the anthropologists because of the presence of particular phenotypic features shared with *H. floresiensis* - which might indicate a possible hybrid origin of these individuals. Notably, these data represent to date the first whole genomes from Indonesia, and the first genetic survey of the Flores pygmies genetic diversity.

Results from the comparison with the Neandertal and Denisova genomes, show that the Flores pygmies, in addition to Neandertal ancestry, harbor a low proportion of Denisova ancestry, which could possibly have been introduced by gene flow from Oceanian populations, after the latter admixed with Denisovans or a Denisova-related group.

Unfortunately, the hot and humid conditions in Flores are not conducive to DNA and attempts at recovering ancient DNA directly from *H. floresiensis* fossils have failed, thus limiting our ability to directly compare our Flores genomes to *H. floresiensis* DNA.

However, using a newly developed computational method for archaic admixture inference that allows to identify traces of introgression from unsequenced hominins, such as *H. floresiensis*, we have found some deeply divergent haplotypes in the genomes of the Flores pygmies. Notably, these putative introgressed haplotypes are inferred to coalesce beyond the modern human and archaic (Neandertal and Denisova) tree – consistent with the hypothesis that they were inherithed through admixture with *H. floresiensis* or a related hominin species.

Taken together these preliminary results, suggest that during Late Pleistocene, Island Southeast Asia was theater to a complex history of interactions among hominin groups. Early hominin species, such as *H. erectus*, *H. floresiensis* and possibly Denisovans, were uncontested masters of the scene, until approximately 40 thousand years ago, when anatomically modern humans came to the scene. Moreover, our findings may be consistent with the recent discovery of traces of gene flow into the Denisova genome, from an unknown hominin that diverged from the modern human lineage between 0.9 and 4 million years ago (Prufer et al., 2014). Indeed, researcher have suggested that this unknown hominin group could be what is known as *H. erectus*, based on the current interpretation of the fossil record (Prufer et al., 2014).

Although its origins remain obscure, the *H. erectus* lineage is thought to have evolved in Africa before 1.89 million years ago (Walker and Leakey, 1993; Spoor et al., 2007) and to have ventured out of Africa into Eurasia (Garcia et al., 2010) and Southeast Asia shortly after. Archeological evidence suggests that these early adventurers reached Flores at least 1 Mya (Brumm et al., 2010), where Pleistocene interglacial high sea levels might have contribuited to the isolation of this archaic hominin group, setting up conditions for the insular evolution of *H. floresiensis*. Indeed the leading theory concerning *H. floresiensis* evolution suggests that this lineage may have experienced a dramatic size reduction from a *H. erectus* ancestor, as a result of insular dwarfism.

Our analysis of the Flores genomes have proved that our method is particularly powerful in detecting traces of archaic admixture, and applicable to areas of the planet where no known fossil ancestor has recoverable ancient DNA. Moreover, our findings have contribuited to unearth a history of gene flow between archaic and modern humans in Island Southeast Asia, one of the most interesting areas for the study of human evolutionary history.

FURTHER APPLICATION OF THE S* STATISTICAL FRAMEWORK

Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. **[Vernot, Tucci et al. 2016, *Science*]**

This chapter is extracted from "Benjamin Vernot, Serena Tucci, Janet Kelso, Joshua G. Schraiber, Aaron B. Wolf,Rachel M. Gittelman, Michael Dannemann, Steffi Grote, Rajiv C. McCoy, Heather Norton, Laura B. Scheinfeldt, David A. Merriwether, George Koki, Jonathan S. Friedlaender, Jon Wakefield, Svante Pääbo and Joshua M. Akey, *Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals*. (2016)
Available at:
http://science.sciencemag.org/content/early/2016/03/16/science.aad9416.abstract

Summary of results

Although sequences inherithed from Neandertals, that persist in the genomes of modern humans, have been identified in Eurasians, no studies have been conducted for populations whose ancestors hybridized with both Neandertals and Denisovans. With the purpose of identifying DNA sequences inherited from multiple archaic hominin ancestors, we developed a new approach and applied it to whole-genome sequences from 1,523 geographically diverse individuals, including 35 new Island Melanesian genomes. Totally, we were able to recover 1.34 Gb and 303 Mb of the Neandertal and Denisovan genome, respectively. We leveraged these maps of archaic sequence to show that Neandertal admixture occurred multiple times in different non-African populations, characterized genomic regions that are significantly depleted of archaic sequence, and identified signatures of adaptive introgression.

Description of the Melanesian samples

We sequenced 35 individuals from 11 locations in the Bismarck Archipelago of Northern Island Melanesia (New Hanover/Lavongai and Saint Mattias/Mussau), Papua New Guinea (Figure 10A; Friedlander et al., 2008) to a median depth of 40x, including a trio to facilitate haplotype reconstruction. Specifically, we selected individuals which belong to populations speaking different non-Austronesian languages and their immediate Austronesian-speaking neighbors in different islands. Our analysis were performed on a subset of 27 unrelated individuals.

Archaeological evidence suggests that modern humans reached parts of Island Melanesia as early as 50,000 - 30,000 years ago (Wickler and Spriggs, 1988; Summerhayes et al., 2007; Leavesley and Chappell, 2004). These small groups remained relatively isolated until approximately 3,300 years ago (Summerhayes et al., 2007), when populations with more complex agriculture and seafaring abilities arrived in the Bismarck Archipelago from the northern coast of New Guinea. This was associated with the spread of Austronesian languages from Taiwan 4,000–5,000 years ago (Blust, 1995; Gray et al., 2009). Almost all Austronesian languages spoken in the Pacific belong to its Oceanic branch. Its ancestor, Proto Oceanic, developed in the Bismarck Archipelago along the north shore of New Britain (Ross et al., 2007), associated with an early phase of the Lapita Cultural Complex.

The non-Austronesian languages spoken in New Guinea and Island Melanesia are thought to be remote descendants of languages spoken by the earlier migrants. Unlike the Austronesian languages for which lexical methods for reconstructing proto-languages are applicable, the relationships of these more diverse languages have been more difficult to reconstruct, since they are an extremely diverse set and include a number of unclassifiable isolates (Ross, 2005). Nevertheless, a recent application of cladistics to certain grammatical features and sound systems suggests that the set of non-Austronesian languages spoken in Island Melanesia are related to one another (Dunn et al., 2005), in what has been called the East Papuan Phylum (Wurm, 1975).

Island Melanesians in the context of worldwide population diversity

To investigate population affinities, we integrated our sequencing data with SNP genotypes for 593,269 autosomal SNPs, from 1,937 individuals spanning 159 worldwide populations, to form the “HO Dataset” (Figure 10A; Lazaridis et al., 2014). After pruning, we performed Principal Component Analysis (PCA). As shown in Figure 10B, the first component, which captures 5.79% of total variation, separates Africa from Eurasia and America (for color codes see Figure 10A and Legend in Figure S3). Populations in Middle East, Arabian Peninsula and East Africa are intermediate between the African cluster and the Eurasian cluster. The second component, which accounts for 4.49% of the variation, separates West Eurasia and South Asia from East Asia and Americas. Finally, our Melanesian individuals form a cluster with other Oceanian populations (Papuans from highland New Guinea, Australians and Bougainville from Nasioi) included in the reference panel.

We then used the unsupervised clustering algorithm ADMIXTURE (Alexander et al., 2009) to infer ancestral clusters in our Melanesian samples and 159 worldwide population of the Human Origin dataset (see SI in Vernot, Tucci et al., 2016). We find that at low values of K, the dominant ancestry component is similar among all Oceanian populations and is shared with Asian populations. At K= 5, populations in Melanesia, along with other populations in Oceania (Papuans from New Guinea, Australians and Bouganville), can be distinguished from Asian populations. For values of K > 5, populations in Oceania are charactherized for the presence of a dominant Oceanian-related ancestry component and a small proportion of ancestry shared with populations in East Asia and Siberia. This East Asian-related patch appears to be absent in Papuans from New Guinea and in the Papuan-speakers Baining in Island Melanesia. These results are highly consistent with the highest proportions (~ 74%) of "Papuan" ancestry we estimated in the Baining and with previous studies (Reich et al., 2011).

Melanesians harbour traces of gene flow from Neandertal and Denisova

To disentangle the archaic ancestry signal, we used a PCA projection analysis where we performed PCA on the Altai Neandertal, the Denisovan, and the chimpanzee genome, included in the merged HO Dataset and projected 1,964 present-day humans onto the plane described by the top two principal components. In Figure 10C we plotted mean values for each of the 11 Melanesian populations and the populations of the global set. Unlike other Eurasian populations which tend to cluster all together, we observed that Oceanians appear to be closer to Denisovan, recapitulating previously reported observations (Reich et al., 2010; Skoglund and Jakobsson, 2011; Qin and Stoneking, 2015). In order to separate the signal of Denisova admixture from the signal deriving from Neandertal admixture, we computed a *f4-ratio* statistics (see Materials and Methods for details). Following previous work (Qin and Stoneking, 2015), we use this *f4-ratio* to estimate Denisova admixture proportions in

Oceanians, using Han Chinese to correct for levels of Neandertal ancestry in Oceanians. We estimated that Melanesians harbor between 1.9% and 3.4% of Denisova ancestry (Z-score ≥ 4; Fig. 10D) and our results are consistent with previous studies (Meyer et al., 2012; Qin and Stoneking, 2015).

Identifying Neandertal and Denisovan introgressed sequences

Having demonstrated that our Melanesian individuals have both Neandertal and Denisovan ancestry, we developed an approach to recover and classify archaic sequence. Briefly, we first identify putative introgressed sequence using the statistic S^*, which does not use information from an archaic reference genome (Plagnol and Wall, 2006; Vernot and Akey, 2014) and then refine this set by comparing significant S^* haplotypes to the Neandertal and Denisovan genomes and testing whether they match more than expected by chance. Variation in neutral divergence between archaic groups across loci and incomplete lineage sorting complicate classification of archaic haplotypes as Neandertal or Denisovan. To address this issue, we developed a likelihood method that operates on the bivariate distribution of Neandertal and Denisovan match *p-values*. This framework estimates the proportion of Neandertal, Denisovan, and null sequence in the set of S^* significant haplotypes, identifies archaic haplotypes at a desired false discovery rate (FDR), and probabilistically categorizes them as Neandertal, Denisovan, or ambiguous (i.e., Neandertal or Denisovan status cannot be confidently distinguished).

In addition to our Melanesian samples, we also applied our method to whole-genome sequences from 1,496 geographically diverse individuals studied as part of the 1000 Genomes Project (1000 Genomes Project Consortium, et al., 2015). We evaluated our approach through coalescent simulations and by analyzing African populations. Archaic match *p-values* calculated from null coalescent simulations without archaic admixture and significant S^* sequences in African individuals show similar distributions (Figure 11A), consistent with little to no Neandertal or Denisovan ancestry in most African populations. Notably, Luhya and Gambians do show evidence of having some Neanderthal ancestry, most likely inherited indirectly through recent admixture with non-Africans (Sikora et al., 2014). In Europeans, we see a strong skew of Neandertal, but not Denisovan, match *p-values* toward zero (Figure 11A). In contrast, Melanesians exhibit a marked skew of both Neandertal and Denisovan match *p-values* toward zero (Figure 11A). In aggregate, across all 1,523 non-African individuals analyzed, we recovered 1,340Mb and 304Mb of the Neandertal and Denisovan genomes, respectively, at a FDR = 5%. Melanesian individuals have on average 104Mb of archaic sequence per individual (48.9Mb, 42.9Mb, and 12.2Mb of Neandertal, Denisovan, and ambiguous sequence, respectively; Figure 11B). In contrast, we only call between 0.026Mb (in Esan) to 0.5Mb (in Luhya) of sequence per individual as archaic in Africans, highlighting that our method and error rates are well calibrated. The higher levels of archaic ancestry in Melanesians results in an average of 20 compound homozygous archaic loci per individual, with one Neandertal and one Denisovan haplotype (Figure 11C).

In other non-Africans, we identify on average 65.0Mb, 55.2Mb, and 51.2Mb of archaic sequence in East Asians, South Asians, and Europeans, respectively (Figure 11B). Virtually all of the archaic sequence in these populations is Neandertal in origin, although a small fraction (<1%) of introgressed sequences in East and South Asians are predicted to be Denisovan.

Neandertal admixture occurred multiple times in different non-African populations

We developed a method to test whether two populations have shared or unique admixture histories based on patterns of reciprocal sharing of Neandertal sequence among individuals and validated the expected behavior of our method in simulated data (Figure 12A), and confirm previous observations of an additional admixture event unique to East Asians (Vernot and Akey, 2014; Vernot and Akey, 2015; Kim and Lohmueller, 2015; Figure 12B). We find evidence for an additional pulse of Neandertal admixture in Europeans, East Asians, and South Asians compared to Melanesians (Figure 12B), which is robust to different statistical thresholds used to call Neandertal and Denisovan sequence and determining significance. Conversely, we find no evidence of differences in admixture histories between Europeans and South Asians. Collectively, these data suggest Neandertal admixture occurred at least three distinct times in modern human history (Figure 12C). Although most South Asian populations show shared histories of archaic admixture, we find significant evidence of differential Neandertal admixture between some European and East Asian populations.

Archaic deserts

The density of surviving Neandertal sequence across the genome is heterogeneous (Vernot and Akey, 2014), and regions that are strongly depleted of Neandertal ancestry may represent loci where archaic sequence was deleterious in hybrid individuals and purged from the population. To quantify how unusual Neandertal depleted regions are under neutral models, we performed coalescent simulations, focusing on individuals of European and East Asian ancestry whose demographic histories are known in most detail. Depletions of Neandertal sequence that extend ≥8 Mb are significantly enriched in the observed compared to simulated data (permutation pvalue < 0.01; Figure 13A). Neandertal depleted regions that span at least 8Mb are also significantly depleted of Neandertal sequence in South Asians and Melanesians (KS test, *p-value* $< 10^{-15}$). We find significantly more overlap in regions depleted of Neandertal and Denisovan lineages than expected by chance (permutation *p-value* = 0.0008), consistent with recurrent selection against deleterious archaic sequence. Indeed, deserts of archaic sequence tend to exhibit higher levels of background selection. Regions depleted of archaic lineages are significantly enriched for genes expressed in specific brain regions, particularly in the developing cortex and adult striatum (permutation *p-value* < 0.05). A large region depleted of archaic sequence spans 11 Mb on chromosome 7 and contains the *FOXP2* gene (Figure 13B), which has been associated with speech and language (Maricic et al., 2013). This region is also significantly enriched for genes associated with autism spectrum disorders (Fisher's exact text, *p-value* = 0.008). Although our data shows that large regions depleted of archaic ancestry are inconsistent with neutral evolution, mechanisms other than selection, such as structural variation, could also contribute to the appearance of archaic deserts and thus additional work is necessary to fully understand the origins of such regions.

Signatures of adaptive introgression in the Melanesian genomes

We identified putative adaptively introgressed sequence in Melanesians by identifying archaic haplotypes at unusually high frequencies as determined by coalescent simulations under a wide variety of neutral demographic models. At a frequency threshold of 0.56, corresponding to the 99th percentile of simulated data, we identified 21 independent candidate regions for adaptive introgression (Figure 13C). Fourteen are of Neanderthal origin, three are Denisovan, three are ambiguous, and one segregates both Neanderthal and Denisovan haplotypes. Six regions do not contain any protein-coding genes, and seven high frequency archaic haplotypes span only a single gene. High frequency archaic haplotypes overlap several metabolism related genes, such as *GCG* (a hormone that increases blood glucose levels) and *PLPP1* (a membrane protein involved in lipid metabolism). Moreover, five regions either span or are adjacent to immune related genes, including a haplotype encompassing *GBP4* and *GBP7* (Figure 13D), which are induced by interferon as part of the innate immune response.

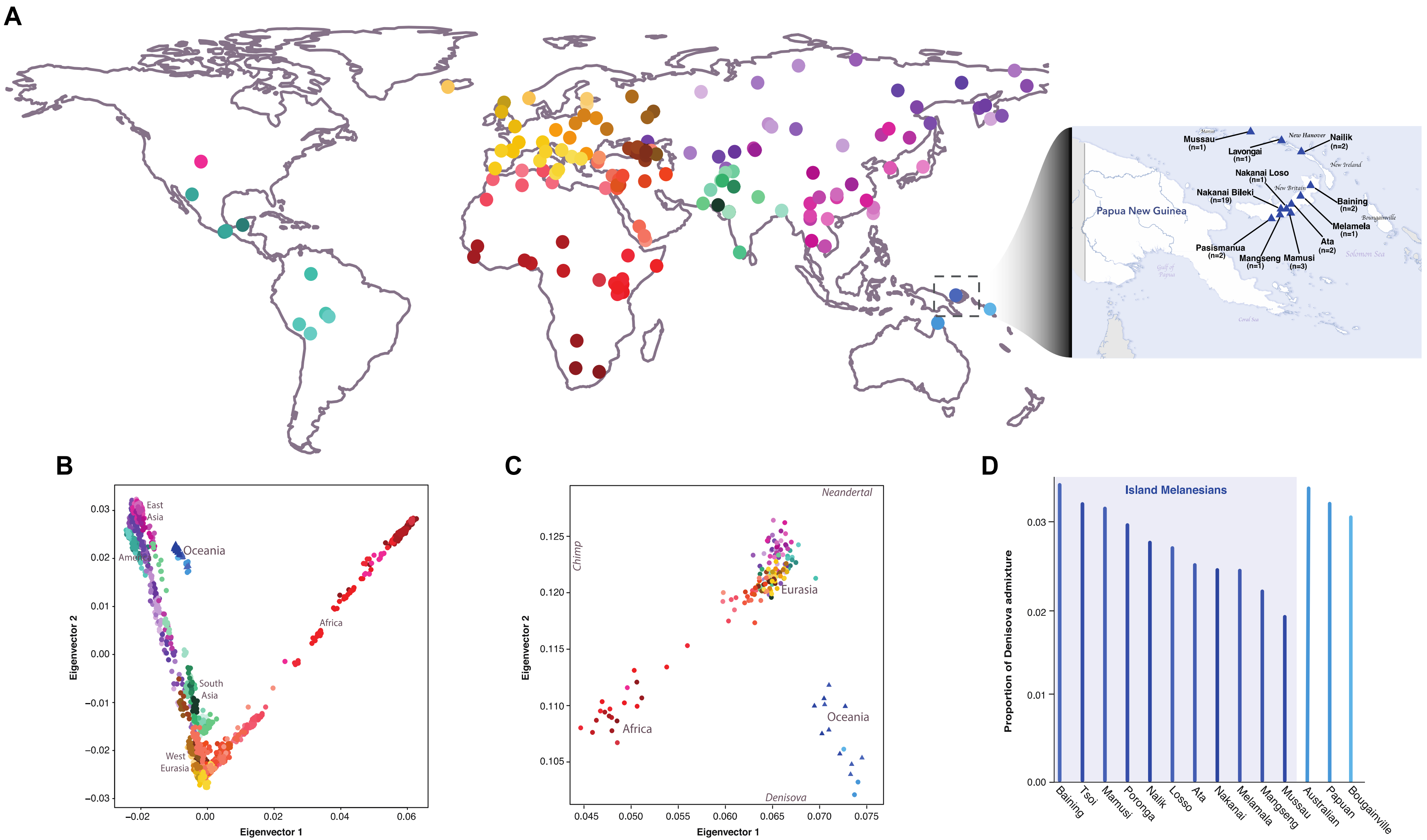

Figure 10. Melanesian genomic variation in a global context. A) Locations of the 159 geographically diverse populations studied. Information on the Melanesian individuals sequenced (blue triangles) is diversity. C) Modern human variation projected onto the top two eigenvectors defined by PCA of the Altai Neandertal, Denisova, and chimpanzee genome. Population means were plotted for each of the 11 Melanesian populations and each population of the global dataset. D) Estimates of Denisova ancestryin Oceanic populations estimated from an f4 statistic. The 11 Melanesian populations are highlighted by the lightblue box.

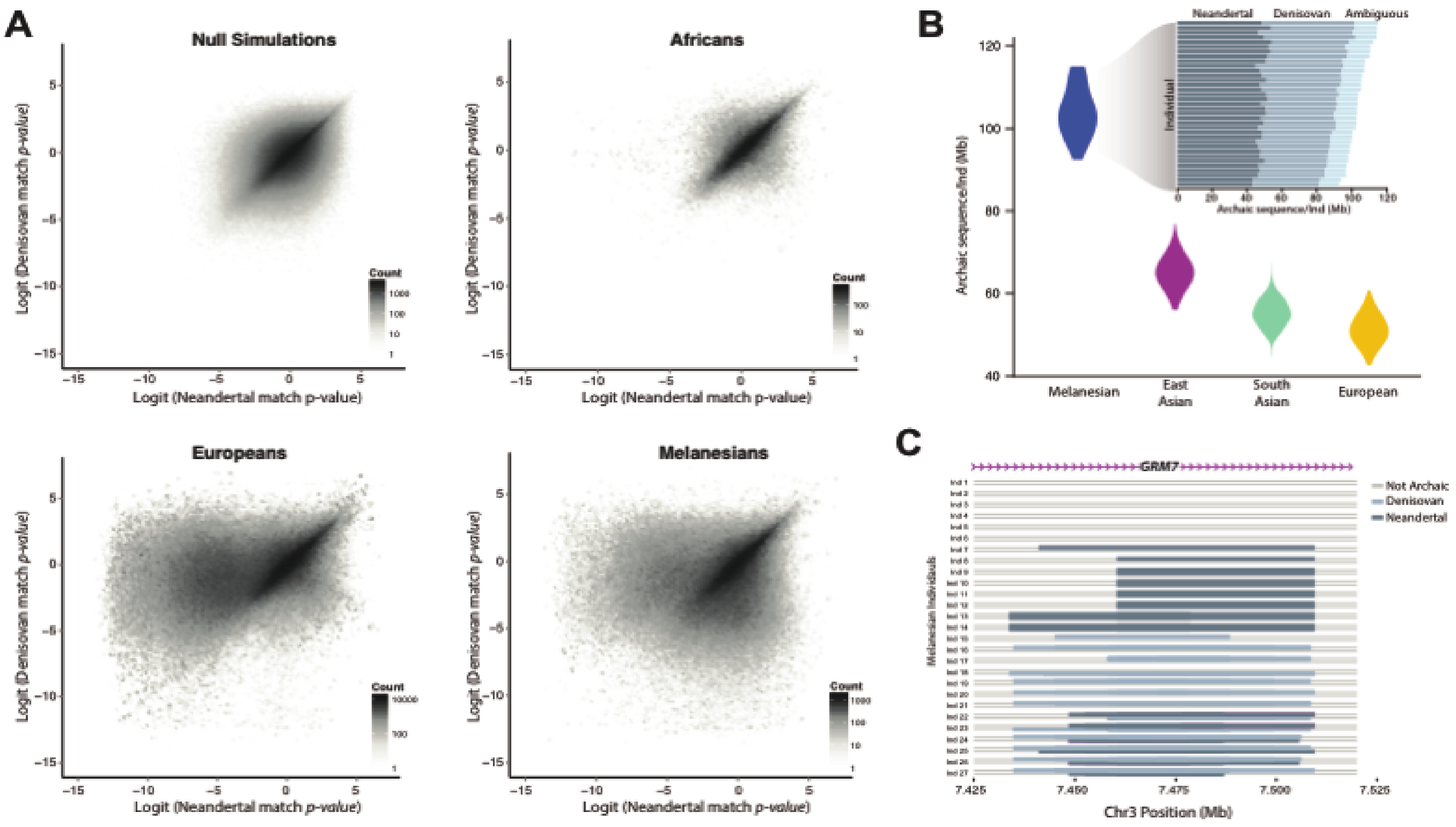

Figure 11. Identifying Neandertal and Denisovan sequences in modern human genomes. A) Bivariate archaic match p-value distributions for simulations of non-introgressed sequence, Esan in Nigeria, Europeans, and Melanesians. Null simulations and Esan show no skew in Neandertal or Denisovan match p-values toward zero, Europeans show only a skew of Neandertal match p-values toward zero, and Melanesians exhibit both Neandertal and Denisovan match p-values skewed toward zero. B) Amount of archaic introgressed sequence identified in each population. Inset, amount of Neandertal, Denisovan, and ambiguous (Neandertal or Denisovan) introgressed sequence for each Melanesian individual. C) Schematic representation of introgressed haplotypes in an intronic portion of the GRM7 locus in Melanesian individuals illustrating mosaic patterns of archaic ancestry.

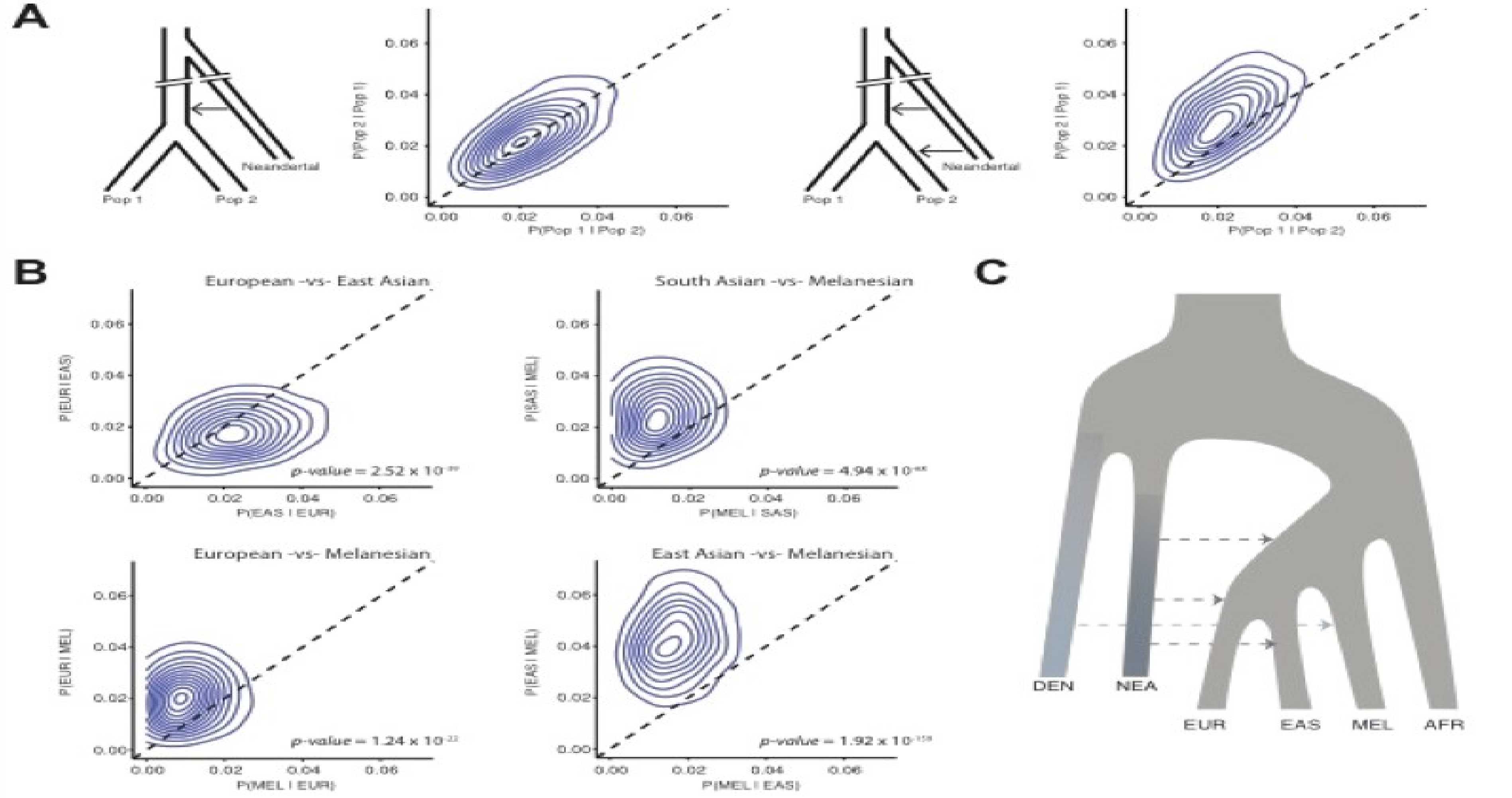

Figure 12. Identifying shared and unique pulses of Neandertal admixture among human populations. A) Schematics of two simulated introgression models, and patterns of reciprocal match probabilities. Contour plots are fit to the scatterplot of reciprocal match probabilities calculated from analyzing all pairwise combinations of individuals between two populations. Left, gene flow occurs into the common ancestor of Population 1 and Population 2, and reciprocal match probabilities fall along the diagonal as predicted by theory (Binomial test, p-value > 0.05). Right, Population 2 receives additional admixture shifting reciprocal match probabilities above the diagonal (Binomial test, p-value < 0.05). B) Reciprocal match probabilities of Neandertal sequences in modern human populations, consistent with additional Neandertal admixture into East Asians versus Europeans, and into Europeans, East Asians and South Asians versus Melanesians. C) Schematic of admixture history consistent with the data.

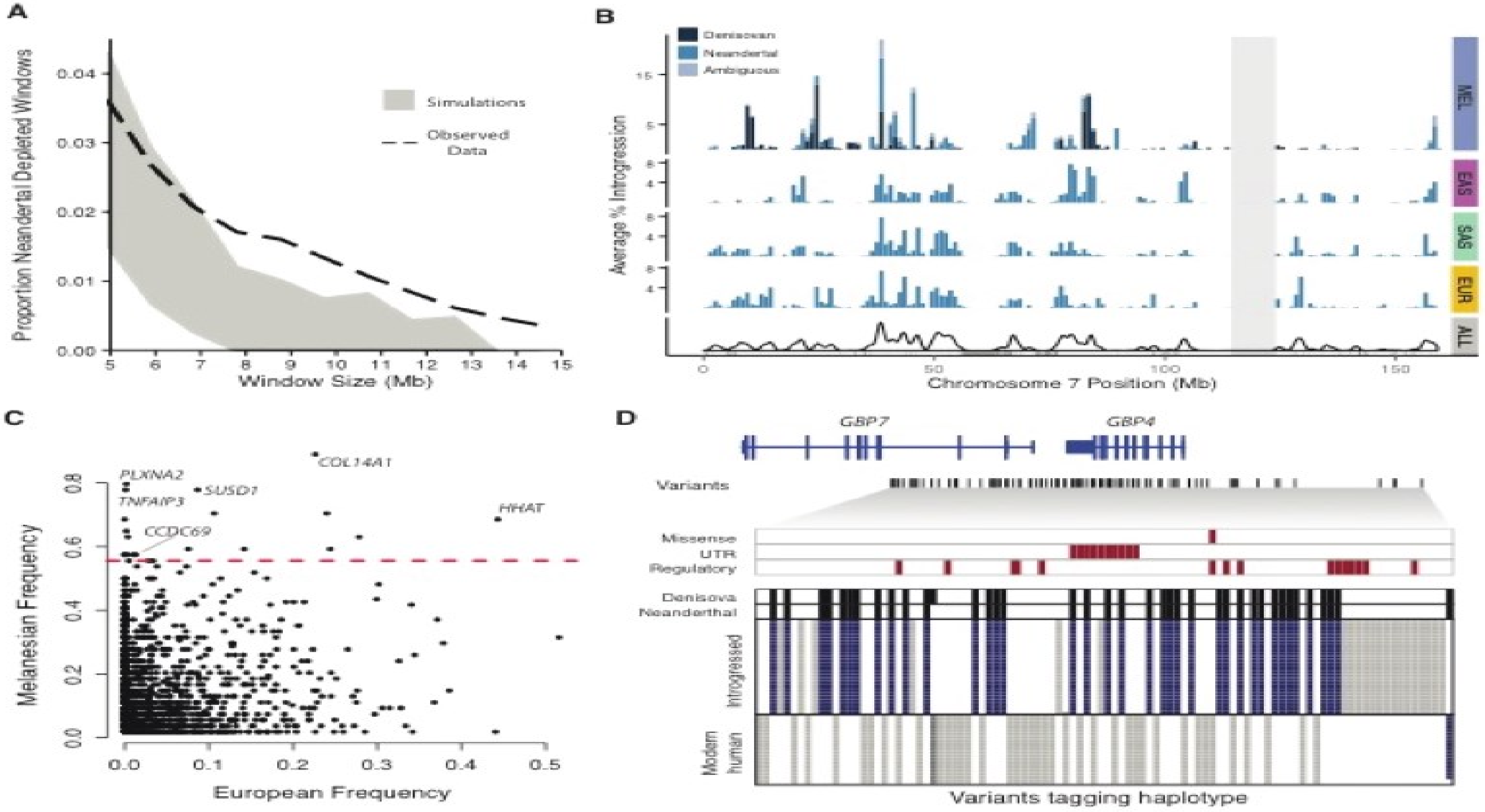

Figure 13. Maps of archaic admixture reveal signatures of purifying and positive selection. A) Proportion of windows significantly depleted of Neandertal introgression in Europeans and East Asians (dashed line) versus what is expected in neutral demographic models (95% CI in gray). B) Distribution of Neandertal and Denisovan sequence across chromosome 7 in Melanesians (MEL), East Asians (EAS), South Asians (SAS), Europeans (EUR), and summed across all populations (TOT). Masked regions are shown as grey vertical lines. An 11.1 Mb region significantly depleted of Denisovan and Neandertal ancestry in all populations is shown in light gray. C) The frequency of archaic haplotypes in Melanesians versus Europeans. The red line indicates the 99th percentile defined by neutral coalescent simulations. Notable genes are labeled. D) Visual representation of a high frequency haplotype encompassing GBP4 and GBP7. Rows indicate individual haplotypes and columns denote variants that tag the introgressed haplotype. Alleles are colored according to whether they are ancestral (white), derived variants that match both archaic genomes (blue), derived variants that match one archaic genome (dark grey), or are derived but do not match either archaic genome (light gray). Archaic sequences are represented above, with black denoting derived variants. Missense, UTR, and putative regulatory variants , are highlighted with red boxes.

CONCLUSION AND FUTURE PERSPECTIVES

In this thesis I have shown that substantial amounts of archaic sequences can now be robustly identified in the genomes of present day human populations, even in the absence of a archaic reference genome, allowing new interesting insights into our recent evolutionary past.

As whole-genome sequencing data from worldwide populations continues to accumulate, a nearly complete map of surviving archaic lineages may be obtaneid soon. Further questions remain, such as the functional and phenotypic consequences of introgressed sequences. Also, key demographic events need to be investigated more in detail, in order to draw demographic models more consistent with the patterns of introgression seen in present-day populations.

Ultimately, this work shows how little is still known about the complex history of hominin interactions, the species involved, and importantly - the legacy that these species might have left to the contemporary human gene pool.

SUPPLEMENTARY TABLES

Table S1. Samples collected in this study. Asterisks (*) denote samples that were later selected for whole genome sequencing.

Sample�ID	Sex	Stature(cm)	Clan
RPS001*	M	135	Kina
RPS002	M	175	Nangka
RPS003	M	152	Taga
RPS004*	M	145	Ntala
RPS005	F	138	Akel
RPS006	M	155	Taga
RPS007	M	150	Ntala
RPS008	M	155	Taga
RPS009*	M	139	Ntala
RPS010	M	153	Taga
RPS011	M	148	Tuke
RPS012	M	155	Lao
RPS013*	M	130	Tuke
RPS014	M	150	Ntala
RPS015*	F	132	Racang
RPS016	M	153	Taga
RPS017	M	142	Tuke
RPS018*	M	136	Taga
RPS019	M	153	Tuke
RPS020*	M	143	Taga
RPS021	F	137	Timur
RPS022	M	147	Taga
RPS023	F	146	Taga
RPS024	F	145	Likung
RPS025	F	140	Kina�
RPS026	F	140	Ntala
RPS027*	F	138	Kina�
RPS028*	F	132	Wae�Kina
RPS029	M	143	Leda
RPS030	F	152	Taga
RPS031*	F	133	Tuke
RPS032	F	150	Wangpu

Table S2. Pan-Asian HUGO dataset used in the first part of the workflow in order to select individuals for whole-genome sequencing.

PANASIA HUGO CONSORTIUM DATASET

Pop Code	Source	Language	Language family	n
AXAM	Taiwan_Affymetrix	Ami	Austronesian	10
AXAT	Taiwan_Affymetrix	Atayal	Austronesian	10
AXME	Melanesian_Affymetrix	Naasioi	Indo-Pacific (East Papuan)	4
CHB	China_HapMap	Chinese	Sino-Tibetan	45
CNCC	China	Zhuang,	Northern Tai-Kedai	24
CNGA	China	Cantonese	Sino-Tibetan	30
CNHM	China	Hmong	Miao-Yao	19
CNJI	China	Jiamao	Tai-Kedai	31
CNJN	China	Jinuo	Sino-Tibetan (Loloish)	26
CNSH	China	Mandarin	Sino-Tibetan	21
CNUG	China	Uyghur	Altaic	26
CNWA	China	Wa	Austro-Asiatic	48
IDAL	Indonesia	Alorese	Austronesian	19
IDDY	Indonesia	Dayak	Austronesian	12
IDJA	Indonesia	Javanese	Austronesian	34
IDJV	Indonesia	Javanese	Austronesian	19
IDKR	Indonesia	Karo	Batak Karo Austronesian	17
IDLA	Indonesia	Lamaholot	Austronesian	20
IDLE	Indonesia	Lembata	Austronesian	19
IDML	Indonesia	Malay	Austronesian	11
IDMT	Indonesia	Mentawai	Austronesian	15
IDRA	Indonesia	Manggarai	Austronesian	9
IDSB	Indonesia	Kambera	Austronesian	20
IDSO	Indonesia	Manggarai	Austronesian	18
IDSU	Indonesia	Sunda	Austronesian	25
IDTB	Indonesia	Toba	Batak Toba Austranesian	20
IDTR	Indonesia	Toraja	Austronesian	20
INDR	India	Telugu,	Kannada Dravidian	23
INEL	India	Bengali	Indo-European	10
INIL	India	Hindi	Indo-European	12
INNI	India	Pahari	Indo-European	20
INNL	India	Hindi	Indo-European	13
INSP	India	Hindi	Indo-European	20
INTB	India	Spiti	Sino-Tibetan	23
INWI	India	Bhili	Indo-European	23
INWL	India	Marathi	Indo-European	12
JPML	Japan	Japanese	Altaic	71
JPRK	Japan	Okinawan	Altaic	43
JPT	Japan_HapMap	Japanese	Altaic	44
KRKR	Korea	Korean	Altaic	90
MYBD	Malaysia	Jagoi	Austronesian	41
MYJH	Malaysia	Jehai	Austro-Asiatic	34
MYKN	Malaysia	Malay	Austronesian	15
MYKS	Malaysia	Kensiu	Austro-Asiatic	22
MYMN	Malaysia	Malay	Austronesian	16
MYTM	Malaysia	Temuan	Austronesian	30
PIAE	Philippines	Ayta	Austronesian	8
PIAG	Philippines	Agta	Austronesian	8
PIAT	Philippines	Ati	Austronesian	22
PIIR	Philippines	Iraya	Austronesian	9
PIMA	Philippines	Manobo	Austronesian	17
PIMW	Philippines	Mamanwa	Austronesian	17
PIUB	Philippines	Ilocano	Austronesian	20
PIUI	Philippines	Visaya, Chabakano	Austranesian	20
PIUN	Philippines	Tagalog	Austronesian	19
SGCH	Singapore	Mandarin	Sino-Tibetan	30
SGID	Singapore	Tamil	Dravidian	30
SGML	Singapore	Malay	Austronesian	30
THHM	Thailand	Hmong	Hmong-Mien	19
THKA	Thailand	Karen	Sino-Tibetan	20
THLW	Thailand	Lawa	Austro-Asiatic	16
THMA	Thailand	Mlabri	Austro-Asiatic	10
THMO	Thailand	Mon	Austro-Asiatic	18
THPL	Thailand	Paluang	Austro-Asiatic	17
THPP	Thailand	Blang	Austro-Asiatic	17
THTK	Thailand	Tai Khuen	Tai-Kadai	17
THTL	Thailand	Lue	Tai-Kadai	18
THTN	Thailand	Mal	Austro-Asiatic	13
THTU	Thailand	Tai Yuan	Tai-Kadai	20
THTY	Thailand	Tai Yong	Tai-Kadai	18
THYA	Thailand	Iu Mien	Hmong-Mien	17
TWHA	Taiwan	Hakka	Sino-Tibetan	48
TWHB	Taiwan	Minnan	Sino-Tibetan	32

Table S3. Sequencing coverage depth statistics for 10 Flores samples. Individual "RPS020" was removed in analyses of unrelated individuals.

Sample ID	Mean depth (genome)	Median depth (Autosomes)	99.5th Percentile
RPS001	31.1	33	62
RPS004	32.4	34	69
RPS009	33.3	34	70
RPS013	37.4	38	75
RPS015	36.6	38	71
RPS018	37.8	40	75
RPS020	42.2	44	82
RPS027	32.2	34	67
RPS028	32.5	34	66
RPS031	47.5	49	92

Table S4. Human Origina Dataset “HO Dataset” used in the PCA, ADMIXTURE, inbreeding analysis and formal tests for admixture.

Population	Latitude	Longitude	n	Region
AA	39.7	-105	10	Africa
Algerian	36.8	3	6	Africa
BantuKenya	-3	37	6	Africa
BantuSA	-29	29	8	Africa
Biaka	4	17	20	Africa
Datog	-3.3	35.7	3	Africa
Egyptian	31	31.2	18	Africa
Esan	6.5	6	8	Africa
Ethiopian_Jew	9	38.7	7	Africa
Gambian	13.4	16.7	6	Africa
Hadza	-3.6	35.1	5	Africa
Ju_hoan_Nort	-18.9	21.5	5	Africa
Khomani	-27.8	21.1	11	Africa
Kikuyu	-0.4	36.9	4	Africa
Libyan_Jew	32.9	13.2	9	Africa
Luhya	1.3	36.8	8	Africa
Luo	-0.1	34.3	8	Africa
Mandenka	12	-12	17	Africa
Masai	-1.5	35.2	12	Africa
Mbuti	1	29	10	Africa
Mende	8.5	-13.2	8	Africa
Moroccan_Jev	34	-6.8	6	Africa
Mozabite	32	3	21	Africa
Saharawi	27.3	-8.9	6	Africa
Somali	5.6	48.3	13	Africa
Tunisian	36.8	10.2	8	Africa
Tunisian_Jew	36.8	10.2	7	Africa
Yoruba	7.4	3.9	70	Africa
Bolivian	-16.5	-68.2	7	America
Karitiana	-10	-63	12	America
Mayan	19	-91	18	America
Mixe	17	-96.6	10	America
Mixtec	16.7	-97.2	10	America
Piapoco	3	-68	4	America
Pima	29	-108	14	America
Quechua	-13.5	-72	5	America
Surui	-11	-62	8	America
Zapotec	17	-96.5	10	America
Aleut	53.6	160.8	7	Central_Asia
Altaian	51.9	86	7	Central_Asia
Chukchi	69.5	168.8	23	Central_Asia
Dolgan	73	115.4	3	Central_Asia
Eskimo	64.5	172.9	22	Central_Asia
Even	57.5	135.9	9	Central_Asia
Itelmen	57.2	156.9	6	Central_Asia
Kalmyk	46.2	45.3	10	Central_Asia
Koryak	58.1	159	9	Central_Asia
Kyrgyz	42.9	74.6	9	Central_Asia
Mansi	62.5	63.3	8	Central_Asia
Mongola	45	111	6	Central_Asia
Nganasan	71.1	96.1	11	Central_Asia
Selkup	65.5	82.3	10	Central_Asia
Tajik_Pomiri	37.4	71.7	8	Central_Asia
Tlingit	54.7	164.5	4	Central_Asia
Tubalar	51.1	87	21	Central_Asia
Turkmen	42.5	59.6	7	Central_Asia
Tuvinian	50.3	95.2	10	Central_Asia
Ulchi	52.2	140.4	25	Central_Asia
Uzbek	41.3	69.2	10	Central_Asia
Yakut	63	129.5	20	Central_Asia
Yukagir	65.5	151	18	Central_Asia
Ami	22.8	121.2	10	East_Asia
Atayal	24.6	121.3	9	East_Asia
Cambodian	12	105	8	East_Asia
Dai	21	100	10	East_Asia
Daur	48.5	124	9	East_Asia
Han	32.3	114	33	East_Asia
Han_NChina	32.3	114	10	East_Asia
Hezhen	47.5	133.5	8	East_Asia
Japanese	38	138	29	East_Asia
Kinh	21	105.9	7	East_Asia
Korean	37.6	127	6	East_Asia
Lahu	22	100	8	East_Asia
Miao	28	109	10	East_Asia
Naxi	26	100	9	East_Asia
Oroqen	50.4	126.5	9	East_Asia
She	27	119	10	East_Asia
Thai	13.8	100.5	8	East_Asia
Tu	36	101	10	East_Asia
Tujia	29	109	10	East_Asia
Uygur	44	81	10	East_Asia
Xibo	43.5	81.5	7	East_Asia

Yi	28	103	10	East_Asia
Flores	*-8.31*	*120.26*	*9*	*ISEA*
Australian	-13	143	3	Oceania
Bougainville	-6	155	10	Oceania
Papuan	-4	143	14	Oceania
Balochi	30.5	66.5	20	South_Asia
Bengali	23.7	90.4	7	South_Asia
Brahui	30.5	66.5	21	South_Asia
Burusho	36.5	74	23	South_Asia
Cochin_Jew	10	76.3	5	South_Asia
GujaratiA	23.2	72.7	5	South_Asia
GujaratiB	23.2	72.7	5	South_Asia
GujaratiC	23.2	72.7	5	South_Asia
GujaratiD	23.2	72.7	5	South_Asia
Hazara	33.5	70	14	South_Asia
Kalash	36	71.5	18	South_Asia
Kusunda	28.1	82.5	7	South_Asia
Makrani	26	64	20	South_Asia
Pathan	33.5	70.5	19	South_Asia
Punjabi	31.5	74.3	8	South_Asia
Sindhi	25.5	69	18	South_Asia
Abkhasian	43	41	9	West_Eurasia
Adygei	44	39	16	West_Eurasia
Albanian	41.3	19.8	6	West_Eurasia
Armenian	40.2	44.5	10	West_Eurasia
Ashkenazi_Jev	52.2	21	7	West_Eurasia
Balkar	43.5	43.6	10	West_Eurasia
Basque	43	0	29	West_Eurasia
BedouinA	31	35	25	West_Eurasia
BedouinB	31	35	19	West_Eurasia
Belarusian	53.9	28	10	West_Eurasia
Bergamo	46	10	12	West_Eurasia
Bulgarian	42.2	24.7	10	West_Eurasia
Chechen	43.3	45.7	9	West_Eurasia
Chuvash	56.1	47.3	10	West_Eurasia
Croatian	43.5	16.4	10	West_Eurasia
Cypriot	35.1	33.4	8	West_Eurasia
Czech	50.1	14.4	10	West_Eurasia
Druze	32	35	39	West_Eurasia
English	51.2	0.7	10	West_Eurasia
Estonian	58.5	24.9	10	West_Eurasia
Finnish	60.2	24.9	7	West_Eurasia
French	46	2	25	West_Eurasia
French_South	43.4	-0.6	7	West_Eurasia
Georgian	42.5	41.9	10	West_Eurasia
Georgian_Jew	41.7	44.8	7	West_Eurasia
Greek	40.6	22.9	20	West_Eurasia
Hungarian	47.5	19.1	20	West_Eurasia
Icelandic	64.1	-21.9	12	West_Eurasia
Iranian	35.6	51.5	8	West_Eurasia
Iranian_Jew	35.7	51.4	9	West_Eurasia
Iraqi_Jew	33.3	44.4	6	West_Eurasia
Jordanian	32.1	35.9	9	West_Eurasia
Kumyk	43.3	46.6	8	West_Eurasia
Lebanese	33.8	35.6	8	West_Eurasia
Lezgin	42.1	48.2	9	West_Eurasia
Lithuanian	54.9	23.9	10	West_Eurasia
Maltese	35.9	14.4	8	West_Eurasia
Mordovian	54.2	45.2	10	West_Eurasia
Nogai	44.4	41.9	9	West_Eurasia
North_Ossetia	43	44.7	10	West_Eurasia
Norwegian	60.4	5.4	11	West_Eurasia
Orcadian	59	-3	13	West_Eurasia
Palestinian	32	35	38	West_Eurasia
Russian	61	40	22	West_Eurasia
Sardinian	40	9	27	West_Eurasia
Saudi	18.5	42.5	8	West_Eurasia
Scottish	56	-3.9	4	West_Eurasia
Sicilian	37.1	15.3	11	West_Eurasia
Spanish	37.4	-6	53	West_Eurasia
Spanish_North	43.3	-4	5	West_Eurasia
Syrian	35.1	36.9	8	West_Eurasia
Turkish	39.6	28.5	56	West_Eurasia
Tuscan	43	11	8	West_Eurasia
Ukrainian	50.3	31.6	9	West_Eurasia
Yemen	14	44.6	6	West_Eurasia
Yemenite_Jew	15.4	44.2	8	West_Eurasia

(continues Table S4)

Table S5. Proportion of "Papuan" ancestry in the Flores individuals measured as a ratio between two *f4 statistics* in the form $f4(Yoruba, Papuan; Han, Flores)/f4(Yoruba, Papuan; Han, Australian)$. We computed a standard error using a weighted block jackknife for each estimated quantity (block size set to 5 cM) using ADMIXTOOLS (Patterson et al 2012).

sample_ID	alpha	SD	Zscore
RPS001	0.355	0.020	17.590
RPS004	0.331	0.021	16.110
RPS009	0.387	0.021	18.166
RPS013	0.319	0.020	15.864
RPS015	0.354	0.020	17.274
RPS018	0.333	0.020	16.528
RPS027	0.373	0.020	18.234
RPS028	0.361	0.021	17.550
RPS031	0.344	0.021	16.499

Table S6. *D statistics* calculated in the for *D(X, Yoruba, Altai Neandertal, Chimp)* to test for the presence of gene flow from the Altai Neandertal in the HO Dataset.

Population	Region	Dstat	Z
AA	Africa	0.0055	3.598
Algerian	Africa	0.021	7.019
BantuKenya	Africa	0.0014	0.767
BantuSA	Africa	0.0053	3.602
Biaka	Africa	0.0026	1.404
Datog	Africa	0.0078	2.942
Egyptian	Africa	0.0207	7.438
Esan	Africa	0.0037	2.52
Ethiopian_Jew	Africa	0.0139	5.425
Gambian	Africa	0.0016	0.896
Hadza	Africa	0.0024	0.791
Ju_hoan_North	Africa	0.019	6.399
Khomani	Africa	0.0171	6.932
Kikuyu	Africa	0.0051	2.245
Libyan_Jew	Africa	0.0262	7.832
Luhya	Africa	1.00E-04	0.039
Luo	Africa	0.0032	1.985
Mandenka	Africa	0.0029	2.247
Masai	Africa	0.0067	3.493
Mbuti	Africa	0.0039	1.513
Mende	Africa	0.0031	2.056
Moroccan_Jew	Africa	0.0222	6.57
Mozabite	Africa	0.0232	8.173
Saharawi	Africa	0.0171	5.497
Somali	Africa	0.0088	4.271
Tunisian	Africa	0.0217	7.421
Tunisian_Jew	Africa	0.0243	7.025
Bolivian	America	0.0296	6.806
Karitiana	America	0.0295	6.55
Mayan	America	0.027	6.594
Mixe	America	0.0291	6.632
Mixtec	America	0.0273	6.341
Piapoco	America	0.0252	5.357
Pima	America	0.0261	5.833
Quechua	America	0.0291	6.664
Surui	America	0.0329	6.566
Zapotec	America	0.0259	5.95
Aleut	Central_Asia	0.0283	7.789
Altaian	Central_Asia	0.0281	7.232
Chukchi	Central_Asia	0.0282	6.833
Dolgan	Central_Asia	0.0249	5.689
Eskimo	Central_Asia	0.0263	6.314
Even	Central_Asia	0.0262	7.157
Itelmen	Central_Asia	0.0266	6.135
Kalmyk	Central_Asia	0.0285	7.49
Koryak	Central_Asia	0.0267	6.215
Kyrgyz	Central_Asia	0.0287	7.482
Mansi	Central_Asia	0.028	7.217
Mongola	Central_Asia	0.0254	6.339
Nganasan	Central_Asia	0.0274	6.578
Selkup	Central_Asia	0.0256	6.723
Tajik_Pomiri	Central_Asia	0.0277	8.331
Tlingit	Central_Asia	0.0274	6.747
Tubalar	Central_Asia	0.0294	7.971
Turkmen	Central_Asia	0.0274	8.148
Tuvinian	Central_Asia	0.0277	7.035
Ulchi	Central_Asia	0.0275	6.877
Uzbek	Central_Asia	0.0288	8.319
Yakut	Central_Asia	0.026	6.634
Yukagir	Central_Asia	0.0283	7.389
Ami	East_Asia	0.0294	7.057
Atayal	East_Asia	0.0298	6.769
Cambodian	East_Asia	0.0271	6.878
Dai	East_Asia	0.0281	6.792
Daur	East_Asia	0.0275	6.853
Han	East_Asia	0.0275	6.898
Han_NChina	East_Asia	0.0294	7.314
Hezhen	East_Asia	0.0254	6.407
Japanese	East_Asia	0.0273	6.819
Kinh	East_Asia	0.0282	6.96
Korean	East_Asia	0.0284	6.832
Lahu	East_Asia	0.024	5.827
Miao	East_Asia	0.0305	7.36
Naxi	East_Asia	0.0274	6.936
Oroqen	East_Asia	0.0268	6.589
She	East_Asia	0.0281	6.776
Thai	East_Asia	0.029	7.388
Tu	East_Asia	0.0261	6.714
Tujia	East_Asia	0.0285	6.871
Uygur	East_Asia	0.0257	7.658
Xibo	East_Asia	0.0282	6.995

Yi	East_Asia	0.029	7.13
Flores	ISEA	0.0306	7.535
Australian	Oceania	0.0384	7.752
Bougainville	Oceania	0.0368	7.594
Papuan	Oceania	0.0369	8.039
Balochi	South_Asia	0.0257	8.346
Bengali	South_Asia	0.0291	8.387
Brahui	South_Asia	0.0257	7.974
Burusho	South_Asia	0.0266	8.131
Cochin_Jew	South_Asia	0.0293	8.315
GujaratiA	South_Asia	0.0257	7.199
GujaratiB	South_Asia	0.0252	7.458
GujaratiC	South_Asia	0.0272	7.643
GujaratiD	South_Asia	0.0273	7.673
Hazara	South_Asia	0.0271	7.791
Kalash	South_Asia	0.0272	7.635
Kusunda	South_Asia	0.0265	6.658
Makrani	South_Asia	0.0257	8.191
Pathan	South_Asia	0.0257	7.819
Punjabi	South_Asia	0.0256	7.354
Sindhi	South_Asia	0.028	8.658
Abkhasian	West_Eurasia	0.0261	7.364
Adygei	West_Eurasia	0.0275	8.378
Albanian	West_Eurasia	0.0291	8.186
Armenian	West_Eurasia	0.0261	7.962
Ashkenazi_Jew	West_Eurasia	0.025	7.292
Balkar	West_Eurasia	0.0273	7.675
Basque	West_Eurasia	0.0263	7.685
BedouinA	West_Eurasia	0.0226	7.845
BedouinB	West_Eurasia	0.0244	7.572
Belarusian	West_Eurasia	0.0294	8.377
Bergamo	West_Eurasia	0.0271	7.904
Bulgarian	West_Eurasia	0.0297	8.459
Chechen	West_Eurasia	0.0271	7.907
Chuvash	West_Eurasia	0.0271	7.725
Croatian	West_Eurasia	0.0264	7.556
Cypriot	West_Eurasia	0.0271	7.917
Czech	West_Eurasia	0.0283	7.963
Druze	West_Eurasia	0.0231	7.444
English	West_Eurasia	0.027	7.668
Estonian	West_Eurasia	0.0284	7.941
Finnish	West_Eurasia	0.0286	7.765
French	West_Eurasia	0.0285	8.479
French_South	West_Eurasia	0.0297	8.309
Georgian	West_Eurasia	0.0295	8.575
Georgian_Jew	West_Eurasia	0.0246	6.878
Greek	West_Eurasia	0.0286	8.569
Hungarian	West_Eurasia	0.0281	8.382
Icelandic	West_Eurasia	0.0272	7.735
Iranian	West_Eurasia	0.0241	7.382
Iranian_Jew	West_Eurasia	0.0262	7.524
Iraqi_Jew	West_Eurasia	0.0264	7.483
Jordanian	West_Eurasia	0.0233	7.417
Kumyk	West_Eurasia	0.0254	7.478
Lebanese	West_Eurasia	0.0231	6.959
Lezgin	West_Eurasia	0.0266	7.659
Lithuanian	West_Eurasia	0.0259	7.147
Maltese	West_Eurasia	0.0273	7.918
Mordovian	West_Eurasia	0.0284	8.103
Nogai	West_Eurasia	0.0269	8.09
North_Ossetian	West_Eurasia	0.0288	8.513
Norwegian	West_Eurasia	0.0276	7.977
Orcadian	West_Eurasia	0.0283	8.009
Palestinian	West_Eurasia	0.0249	8.312
Russian	West_Eurasia	0.0299	8.719
Sardinian	West_Eurasia	0.0267	7.684
Saudi	West_Eurasia	0.0239	7.554
Scottish	West_Eurasia	0.029	7.892
Sicilian	West_Eurasia	0.0269	7.945
Spanish	West_Eurasia	0.027	8.251
Spanish_North	West_Eurasia	0.0293	7.848
Syrian	West_Eurasia	0.024	7.603
Turkish	West_Eurasia	0.0262	8.358
Tuscan	West_Eurasia	0.0282	8.203
Ukrainian	West_Eurasia	0.0292	8.119
Yemen	West_Eurasia	0.0218	7.28
Yemenite_Jew	West_Eurasia	0.025	7.623

(continues Table S6)

Table S7. Differential contribution of Neandertal and Denisova to populations in Asia and Americas, as measured by the *D(Yoruba, X, Altai Neandertal, Denisova)*. We computed a standard error using a weighted block jackknife for each estimated quantity (block size set to 5 cM) using ADMIXTOOLS (Patterson et al., 2012).

Population	Region	Dstat	Z
Karitiana	America	-0.0348	-6.099
Piapoco	America	-0.0351	-5.932
Pima	America	-0.036	-6.597
Quechua	America	-0.0367	-6.995
Mayan	America	-0.037	-7.425
Mixtec	America	-0.0374	-7.132
Bolivian	America	-0.0379	-7.026
Zapotec	America	-0.0381	-7.331
Mixe	America	-0.0399	-7.203
Surui	America	-0.044	-7.163
Selkup	Central_Asia	-0.0361	-7.716
Turkmen	Central_Asia	-0.0363	-8.298
Tajik_Pomiri	Central_Asia	-0.0364	-8.399
Tlingit	Central_Asia	-0.037	-7.36
Tuvinian	Central_Asia	-0.0377	-7.622
Aleut	Central_Asia	-0.0378	-7.946
Mansi	Central_Asia	-0.0382	-7.922
Even	Central_Asia	-0.0384	-8.117
Mongola	Central_Asia	-0.0384	-8.214
Yakut	Central_Asia	-0.0389	-8.013
Uzbek	Central_Asia	-0.0391	-8.856
Kyrgyz	Central_Asia	-0.0392	-8.418
Eskimo	Central_Asia	-0.0393	-7.679
Tubalar	Central_Asia	-0.0395	-8.463
Nganasan	Central_Asia	-0.0398	-7.527
Altaian	Central_Asia	-0.0399	-8.466
Kalmyk	Central_Asia	-0.041	-8.833
Ulchi	Central_Asia	-0.0413	-8.558
Yukagir	Central_Asia	-0.0419	-9.079
Dolgan	Central_Asia	-0.043	-7.769
Koryak	Central_Asia	-0.0431	-8.395
Chukchi	Central_Asia	-0.0433	-8.863
Itelmen	Central_Asia	-0.0443	-8.421
Naxi	East_Asia	-0.0366	-7.471
Lahu	East_Asia	-0.037	-7.358
Tu	East_Asia	-0.0381	-8.081
Uygur	East_Asia	-0.0385	-8.993
Daur	East_Asia	-0.0389	-8.037
Dai	East_Asia	-0.0392	-7.839
Yi	East_Asia	-0.0394	-8.064
Hezhen	East_Asia	-0.0398	-8.164
Han	East_Asia	-0.0402	-8.501
Korean	East_Asia	-0.0403	-8.081
Thai	East_Asia	-0.0403	-8.488
Japanese	East_Asia	-0.0404	-8.433
She	East_Asia	-0.0409	-8.179
Kinh	East_Asia	-0.041	-8.244
Xibo	East_Asia	-0.0413	-8.586
Cambodian	East_Asia	-0.0415	-8.486
Ami	East_Asia	-0.0416	-7.982
Tujia	East_Asia	-0.0417	-8.577
Miao	East_Asia	-0.0418	-8.416
Han_NChina	East_Asia	-0.0424	-8.886
Oroqen	East_Asia	-0.0427	-8.799
Atayal	East_Asia	-0.0441	-8.041
RPS	ISEA	-0.0299	-5.993
Papuan	Oceania	-0.0036	-0.571
Australian	Oceania	-0.0052	-0.773
Bougainville	Oceania	-0.0126	-2.026
GujaratiB	South_Asia	-0.0343	-7.686
Sindhi	South_Asia	-0.0352	-8.451
Makrani	South_Asia	-0.0355	-8.829
Pathan	South_Asia	-0.0355	-8.525
Punjabi	South_Asia	-0.0355	-7.574
Burusho	South_Asia	-0.0359	-8.442
Balochi	South_Asia	-0.0364	-8.747

Brahui	South_Asia	-0.0364	-8.732
GujaratiD	South_Asia	-0.037	-8.052
Kalash	South_Asia	-0.0373	-8.162
Hazara	South_Asia	-0.0374	-8.587
GujaratiA	South_Asia	-0.0375	-8.307
Kusunda	South_Asia	-0.0379	-7.591
Bengali	South_Asia	-0.0384	-8.793
GujaratiC	South_Asia	-0.0384	-8.487
Cochin_Jew	South_Asia	-0.0389	-8.799
Lebanese	West_Eurasia	-0.0304	-6.969
Saudi	West_Eurasia	-0.032	-7.851
BedouinA	West_Eurasia	-0.0331	-8.925
Syrian	West_Eurasia	-0.0331	-7.987
Lithuanian	West_Eurasia	-0.0342	-7.388
Druze	West_Eurasia	-0.0343	-8.276
Yemenite_Jew	West_Eurasia	-0.0345	-8.223
Palestinian	West_Eurasia	-0.0348	-9.302
BedouinB	West_Eurasia	-0.0351	-8.502
Chechen	West_Eurasia	-0.0353	-8.03
Kumyk	West_Eurasia	-0.0354	-8.003
English	West_Eurasia	-0.0355	-7.917
Lezgin	West_Eurasia	-0.0356	-8.025
Yemen	West_Eurasia	-0.0358	-8.935
Icelandic	West_Eurasia	-0.0361	-8.029
Cypriot	West_Eurasia	-0.0361	-8.252
Abkhasian	West_Eurasia	-0.0363	-8.19
Iranian_Jew	West_Eurasia	-0.0363	-8.17
Basque	West_Eurasia	-0.0363	-8.401
Armenian	West_Eurasia	-0.0364	-8.77
Georgian_Jew	West_Eurasia	-0.0364	-8.313
Iraqi_Jew	West_Eurasia	-0.0368	-8.089
Turkish	West_Eurasia	-0.0368	-9.113
Sicilian	West_Eurasia	-0.0368	-8.461
Nogai	West_Eurasia	-0.037	-8.661
Iranian	West_Eurasia	-0.0371	-8.69
Balkar	West_Eurasia	-0.0372	-8.264
Bergamo	West_Eurasia	-0.0372	-8.676
Ashkenazi_Jew	West_Eurasia	-0.0372	-8.264
Estonian	West_Eurasia	-0.0372	-8.397
Norwegian	West_Eurasia	-0.0372	-8.146
Jordanian	West_Eurasia	-0.0373	-9.041
Spanish	West_Eurasia	-0.0374	-8.991
Tuscan	West_Eurasia	-0.0377	-8.541
Chuvash	West_Eurasia	-0.0381	-8.583
Ukrainian	West_Eurasia	-0.0382	-8.417
Czech	West_Eurasia	-0.0383	-8.434
Maltese	West_Eurasia	-0.0383	-8.769
Orcadian	West_Eurasia	-0.0384	-8.48
Croatian	West_Eurasia	-0.0385	-8.657
North_Ossetian	West_Eurasia	-0.0385	-8.798
Mordovian	West_Eurasia	-0.0386	-8.519
Sardinian	West_Eurasia	-0.0386	-8.955
Greek	West_Eurasia	-0.0387	-8.995
Finnish	West_Eurasia	-0.0389	-8.379
French	West_Eurasia	-0.0391	-9.252
Russian	West_Eurasia	-0.0393	-8.796
Georgian	West_Eurasia	-0.0394	-8.802
Albanian	West_Eurasia	-0.0395	-8.782
Hungarian	West_Eurasia	-0.0395	-9.167
Belarusian	West_Eurasia	-0.0397	-8.829
Bulgarian	West_Eurasia	-0.0398	-8.763
Adygei	West_Eurasia	-0.0407	-9.337
French_South	West_Eurasia	-0.0411	-9.154
Scottish	West_Eurasia	-0.0419	-8.746
Spanish_North	West_Eurasia	-0.0419	-9.002

(continues Table S7)

Table S8. Denisova ancestry as estimated using a *f4-ratio statistic* in the form *f4(Yoruba, Altai Neandertal; Han Chinese, Flores)/f4(Yoruba, Altai Neandertal; Han Chinese, Denisova)*. We computed a standard error using a weighted block jackknife for each estimated quantity (block size set to 5 cM) using ADMIXTOOLS (Patterson et al., 2012).

sample_ID	alpha	stardard error	zscore
RPS001	0.006389	0.00326	1.96
RPS004	0.009441	0.003259	2.897
RPS009	0.007879	0.003782	2.083
RPS013	0.004027	0.003097	1.3
RPS015	0.009989	0.003609	2.768
RPS018	0.01467	0.003355	4.373
RPS027	0.011839	0.003497	3.386
RPS028	0.015187	0.00344	4.415
RPS031	-0.000791	0.003432	-0.23

SUPPLEMENTARY FIGURES

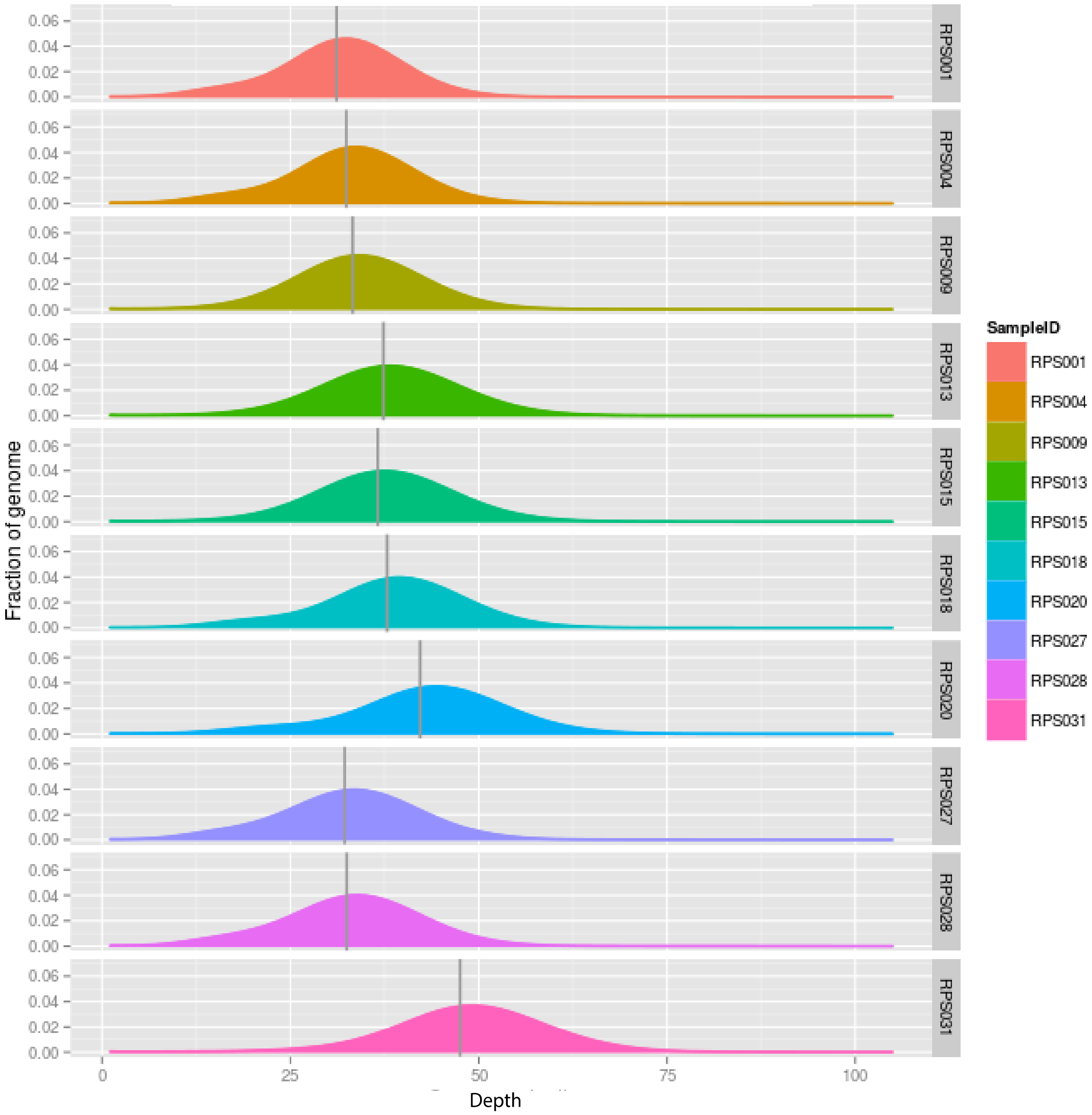

Figure S1. Coverage depth distribution in the Flores genomes. Grey lines denotes average depth estimated for the whole genome.

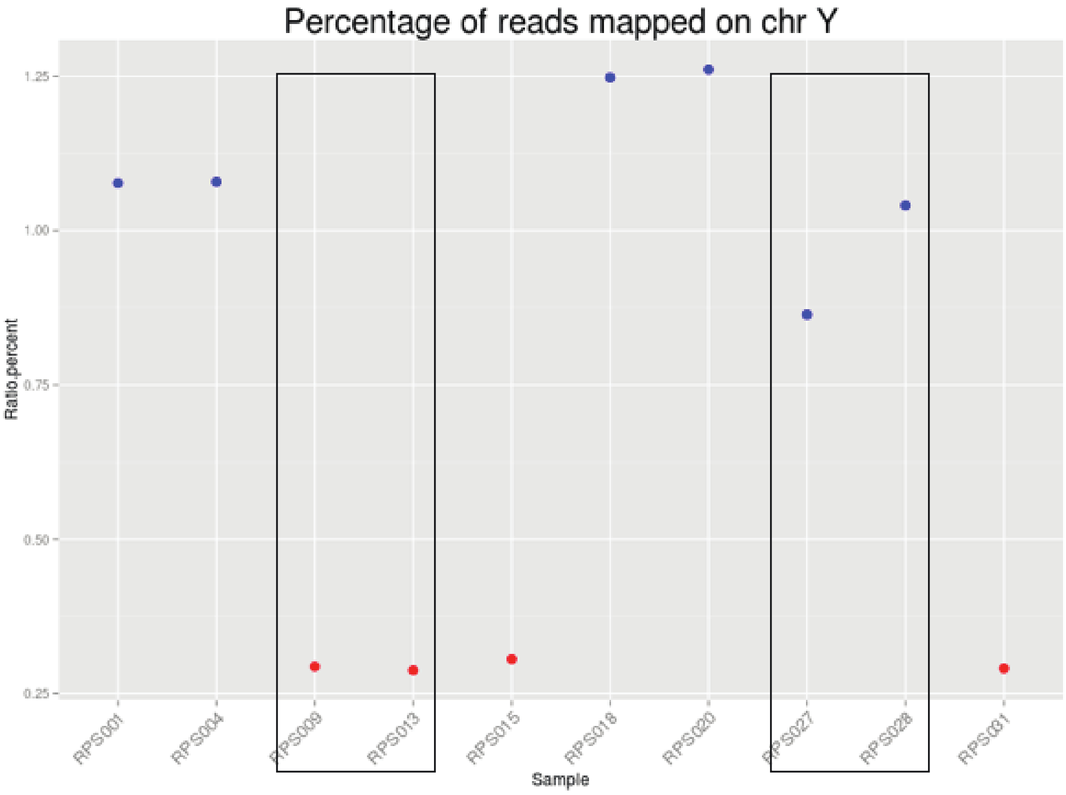

FIGURE S2. Percentage of reads that mapped on chromosome Y. Rectangles show the samples that were mislabeled during the sequencing. Indeed, based on the information collected on the field (Table S1), RPS009 and RPS013 are males, while RPS027 and RPS028 are females, but the number of reads mapped to the Y-chromosome here shows the contrary.

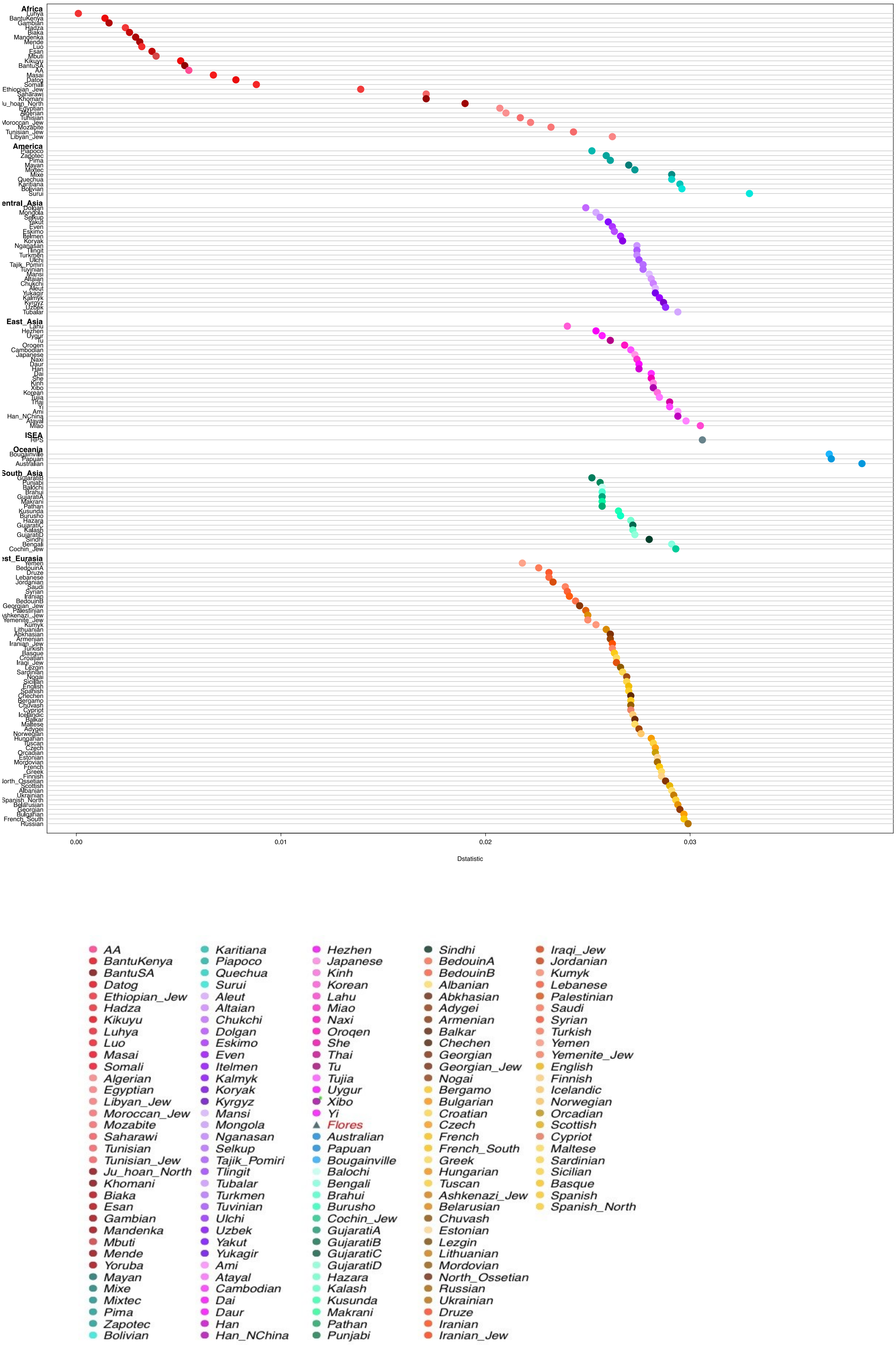

Figure S3. *D statistics* in the form *D(X, Yoruba, Altai Neandertal, Chimp)* calculated for all 159 in the dataset. Populations from Eurasia and Americans show Z scores > 5. The legend is referred to populations included in the HO Dataset.

BIBLIOGRAPHY

Abdulla, M. A., Ahmed, I., Assawamakin, A., Bhak, J., Brahmachari, S. K., Calacal, G. C., … Zilfalil, B. A. (2009). Mapping human genetic diversity in Asia. *Science (New York, N.Y.)*, *326*(5959), 1541–5. doi:10.1126/science.1177074

Abecasis, G. R., Noguchi, E., Heinzmann, A., Traherne, J. A., Bhattacharyya, S., Leaves, N. I., … Cookson, W. O. C. (2016). Extent and Distribution of Linkage Disequilibrium in Three Genomic Regions. *The American Journal of Human Genetics*, *68*(1), 191–197. doi:10.1086/316944

Aghakhanian, F., Yunus, Y., Naidu, R., Jinam, T., Manica, A., Hoh, B. P., & Phipps, M. E. (2015). Unravelling the genetic history of Negritos and Indigenous populations of Southeast Asia. *Genome Biology and Evolution*, *7*(5), 1206–1215. doi:10.1093/gbe/evv065

Aiello, L. C. (2010). Five years of Homo floresiensis. *American Journal of Physical Anthropology*, *142*(2), 167–179. doi:10.1002/ajpa.21255

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, *19*. doi:10.1101/gr.094052.109

Alliel, P. M., Seddiqi, N., Goudou, D., Cifuentes-Diaz, C., Romero, N., Velasco, E., … Périn, J.-P. (2000). Myoneurin, a Novel Member of the BTB/POZ–Zinc Finger Family Highly Expressed in Human Muscle. *Biochemical and Biophysical Research Communications*, *273*(1), 385–391. doi:http://dx.doi.org/10.1006/bbrc.2000.2862

Anderson, C. a, Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, P., & Zondervan, K. T. (2011). Data quality control in genetic case-control association studies. *Nature Protocols*, *5*(9), 1564–1573. doi:10.1038/nprot.2010.116.Data

Argue, D., Morwood, M.J., Sutikna, T., Jatmiko, Saptomo, E.W., (2009). Homo floresiensis: a cladistic analysis. J. Hum. Evol. 57, 623–639.

Armitage, S. J., Jasim, S. A., Marks, A. E., Parker, A. G., Usik, V. I., & Uerpmann, H.-P. (2011). The Southern Route "Out of Africa": Evidence for an Early Expansion of Modern Humans into Arabia. *Science*, *453*(6016), 453–156. doi:10.1594/PANGAEA.755114

Ayub, Q., Mezzavilla, M., Pagani, L., Haber, M., Mohyuddin, A., Khaliq, S., … Tyler-Smith, C. (2015). The kalash genetic isolate: Ancient divergence, drift, and selection. *American Journal of Human Genetics*, *96*(5), 775–783. doi:10.1016/j.ajhg.2015.03.012

Baba, H., Aziz, F., Kaifu, Y., Suwa, G., Kono, R. T., & Jacob, T. (2003). Homo erectus calvarium from the Pleistocene of Java. *Science (New York, N.Y.)*, *299*(5611), 1384–8. doi:10.1126/science.1081676

Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V, Schwartz, S., … Eichler, E. E. (2002). Recent Segmental Duplications in the Human Genome. *Science*, *297*(5583), 1003–1007. Retrieved from http://science.sciencemag.org/content/297/5583/1003.abstract

Barbujani G, Colonna V (2010) Human genome diversity: frequently asked questions. Trends Genet 26:285-296

Barker, G., Barton, H., Bird, M., Daly, P., Datan, I., Dykes, A., … Turney, C. (2007). The "human revolution" in lowland tropical Southeast Asia: the antiquity and behavior of anatomically modern humans at Niah Cave (Sarawak, Borneo). *Journal of Human Evolution*, *52*(3), 243–61. doi:10.1016/j.jhevol.2006.08.011

Bartstra, G.-J., S. Soegondho & A. van der Wijk (1988). Ngandong man: age and artifacts. Journal of Human Evolution 17: 325-337.

Batini, C., Lopes, J., Behar, D. M., Calafell, F., Jorde, L. B., van der Veen, L., … Comas, D. (2011). Insights into the demographic history of African Pygmies from complete mitochondrial genomes. *Molecular Biology and Evolution*, *28*(2), 1099–110. doi:10.1093/molbev/msq294

Blust, R. (1995) The Prehistory of the Austronesian-Speaking Peoples: A View from Language. *Journal of World Prehistory.* 9, 453–510

Bräuer, G., Broeg, H., & Stringer, C. B. (2006). Earliest Upper Paleolithic crania from Mladeč, Czech Republic, and the question of Neanderthal-modern continuity: metrical evidence from the fronto-facial region BT - Neanderthals Revisited: New Approaches and Perspectives. In J.-J. Hublin, K. Harvati, & T. Harrison (Eds.), (pp. 269–279). Dordrecht: Springer Netherlands. doi:10.1007/978-1-4020-5121-0_15

Brown D. E., (1991). Human Universals. McGraw-Hill, New York

Brown, P., Sutikna, T., Morwood, M. J., Soejono, R. P., Jatmiko, Saptomo, E. W., & Due, R. A. (2004). A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia. *Nature*, *431*(7012), 1055–61. doi:10.1038/nature02999

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, *81*(5), 1084–97. doi:10.1086/521987

Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews. Genetics*, *12*(10), 703–714. doi:10.1038/nrg3054

Brumm, A., Jensen, G. M., van den Bergh, G. D., Morwood, M. J., Kurniawan, I., Aziz, F., & Storey, M. (2010). Hominins on Flores, Indonesia, by one million years ago. *Nature*, *464*(7289), 748–52. doi:10.1038/nature08844

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. *Fly*, *6*(2), 80–92. doi:10.4161/fly.19695

Clarkson, C., Smith, M., Marwick, B., Fullagar, R., Wallis, L. A., Faulkner, P., … Florin, S. A. (2015). The archaeology, chronology and stratigraphy of Madjedbebe (Malakunanja II): A site in northern Australia with early occupation. *Journal of Human Evolution*, *83*, 46–64. doi:http://dx.doi.org/10.1016/j.jhevol.2015.03.014

Cloëtta, D., Thomanetz, V., Baranek, C., Lustenberger, R. M., Lin, S., Oliveri, F., … Rüegg, M. A. (2013). Inactivation of mTORC1 in the Developing Brain Causes Microcephaly and Affects Gliogenesis. *The Journal of Neuroscience* , *33* (18), 7799–7810. doi:10.1523/JNEUROSCI.3294-12.2013

Cox, M. P. (2013) in The Encyclopedia of Global Human Migration (Blackwell Publishing Ltd)

Currat, M., & Excoffier, L. (2004). Modern humans did not admix with Neanderthals during their range expansion into Europe. *PLoS Biology*, *2*(12). doi:10.1371/journal.pbio.0020421

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., & DePristo, M. A. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*. doi:10.1093/bioinformatics/btr330

DePristo, M. a, Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–8. doi:10.1038/ng.806

Dubois, E., (1894). Pithecanthropus erectus, eine Menschenähnliche Übergangsform aus Java. Landes Druckerei, Batavia.

Dumas, L. J., O'Bleness, M. S., Davis, J. M., Dickens, C. M., Anderson, N., Keeney, J. G., … Sikela, J. M. (2012). DUF1220-domain copy number implicated in human brain-size pathology and evolution. *American Journal of Human Genetics*, *91*(3), 444–454. doi:10.1016/j.ajhg.2012.07.016

Dunn, M., Terrill, A., Reesink, G. Foley, R. A. ,Levinson, S. C. (2005) Structural

Phylogenetics and the Reconstruction of Ancient Language History. Science. 309, 2072–2075 (2005).

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, *28*(8), 2239–2252. doi:10.1093/molbev/msr048

Enard, W., Przeworski, M., Fisher, S. E., Lai, C. S. L., Wiebe, V., Kitano, T., … Paabo, S. (2002). Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*, *418*(6900), 869–872. Retrieved from http://dx.doi.org/10.1038/nature01025

Eriksson, A., & Manica, A. (2012). Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proc Natl Acad Sci U S A*, *109*. doi:10.1073/pnas.1200567109

Falk, D., Hildebolt, C., Smith, K., Morwood, M. J., Sutikna, T., Brown, P., … Prior. (2005). The brain of LB1, Homo floresiensis. *Science (New York, N.Y.)*, *308*(5719), 242–5. doi:10.1126/science.1109727

Falk, D., Hildebolt, C., Smith, K., Morwood, M. J., Sutikna, T., Jatmiko, … Prior, F. (2007). Brain shape in human microcephalics and Homo floresiensis. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(7), 2513–8. doi:10.1073/pnas.0609185104

Falk, D., Hildebolt, C., Smith, K., Morwood, M. J., Sutikna, T., Jatmiko, … Prior, F. (2009). LB1's virtual endocast, microcephaly, and hominin brain evolution. *Journal of Human Evolution*, *57*(5), 597–607. doi:10.1016/j.jhevol.2008.10.008

Ferring, R., Oms, O., Agustí, J., Berna, F., Nioradze, M., Shelia, T., … Lordkipanidze, D. (2011). Earliest human occupations at Dmanisi (Georgian Caucasus) dated to 1.85–1.78 Ma. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(26), 10432–10436. doi:10.1073/pnas.1106638108

Fleagle, J. G., Assefa, Z., Brown, F. H., & Shea, J. J. (2008). Paleoanthropology of the Kibish Formation, southern Ethiopia: Introduction. *Journal of Human Evolution*, *55*(3), 360–5. doi:10.1016/j.jhevol.2008.05.007

Fu, Q., Hajdinjak, M., Moldovan, O. T., Constantin, S., Mallick, S., Skoglund, P., … Paabo, S. (2015). An early modern human from Romania with a recent Neanderthal ancestor. *Nature*, *advance on*. Retrieved from http://dx.doi.org/10.1038/nature14558

Garcia T., Féraud G., Falguères C., de Lumley H., Perrenoud C., Lordkipanidze D., 2010. Earliest human remains in Eurasia: New 40Ar/39Ar dating of the Dmanisi hominid-bearing levels. *Quaternary Geochronology,* Georgia, 5:443-451.

Ghirotto S., Tassi F., Benazzo A. and Barbujani G. (2011) No evidence of Neandertal admixture in the mitochondrial genomes of early European modern humans and contemporary Europeans. *American Journal of Physical Anthropology* 146:242–252.

Gibson, J., Morton, N. E., & Collins, A. (2006). Extended tracts of homozygosity in outbred human populations. *Human Molecular Genetics*, *15*(5), 789–795. doi:10.1093/hmg/ddi493

Gordon, A. D., Nevell, L., & Wood, B. (2008). The Homo floresiensis cranium (LB1): size, scaling, and early Homo affinities. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(12), 4650–5. doi:10.1073/pnas.0710041105

Gray, R. D., Drummond, A. J., & Greenhill, S. J. (2009). RESEARCH ARTICLE Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement, *323*(January), 479–483.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., … Pääbo, S. (2010). A Draft Sequence of the Neandertal Genome. *Science*, *328*(5979), 710–722. doi:10.1126/science.1188021

Groube, L., Chappell, J., Muke, J., & Price, D. (1986). A 40,000 year-old human occupation site at Huon Peninsula, Papua New Guinea. *Nature*, *324*(6096), 453–5. doi:10.1038/324453a0

Grün, R., & Thorne, A. (1997). Dating the Ngandong Humans. *Science*, *276*(5318), 1575–1576. Retrieved from http://science.sciencemag.org/content/276/5318/1575.abstract

Hammer, M. F., Woerner, A. E., Mendez, F. L., Watkins, J. C., & Wall, J. D. (2011). Genetic evidence for archaic admixture in Africa. *Proc Natl Acad Sci U S A*, *108*. doi:10.1073/pnas.1109300108

Henn, B. M., Gravel, S., Moreno-Estrada, A., Acevedo-Acevedo, S., & Bustamante, C. D. (2010). Fine-scale population structure and the era of next-generation sequencing. *Human Molecular Genetics* , *19* (R2), R221–R226. doi:10.1093/hmg/ddq403

Hennenberg M, Thorne A. (2004). Flores human may be pathological Homo sapiens. Before Farming 4/article 1:2–3.

Hsieh, P., Woerner, A. E., Wall, J. D., Lachance, J., Tishkoff, S. a, Gutenkunst, R. N., & Hammer, M. F. (2016). Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies, 291–300. doi:10.1101/gr.196634.115

Hudson, R. R. (2002) Generating samples under a Wright-Fisher neutral model. Bioinformatics 18, 337-338

Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B. M., Vinckenbosch, N., … Nielsen, R. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, *512*(7513), 194–7. doi:10.1038/nature13408

Huxley TH (1868) On the classification and distribution of the Alectoromorphae and Heteromorphae. Proc Zool Soc London 1868: 296–319.

Indriati, E., & Antón, S. C. (2010). The calvaria of Sangiran 38, Sendangbusik, Sangiran Dome, Java. *Homo : Internationale Zeitschrift Für Die Vergleichende Forschung Am Menschen*, *61*(4), 225–43. doi:10.1016/j.jchb.2010.05.002

Jacob, T., Indriati, E., Soejono, R. P., Hsü, K., Frayer, D. W., Eckhardt, R. B., … Henneberg, M. (2006). Pygmoid Australomelanesian Homo sapiens skeletal remains from Liang Bua, Flores: population affinities and pathological abnormalities. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(36), 13421–6. doi:10.1073/pnas.0605563103

Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, *23*(14), 1801–1806. doi:10.1093/bioinformatics/btm233

Jarvis, J. P., Scheinfeldt, L. B., Soi, S., Lambert, C., Omberg, L., Ferwerda, B., … Tishkoff, S. a. (2012). Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLoS Genetics*, *8*(4). doi:10.1371/journal.pgen.1002641

Jinam, T. a, Phipps, M. E., & Saitou, N. (2013). Admixture patterns and genetic differentiation in negrito groups from West Malaysia estimated from genome-wide SNP data. *Human Biology*, *85*(1-3), 173–88. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24297225

Jolly, C. J., Burrell, a S., Phillips-Conroy, J. E., Bergey, C., & Rogers, J. (2011). Kinda baboons (Papio kindae) and grayfoot chacma baboons (P. ursinus griseipes) hybridize in the Kafue river valley, Zambia. *American Journal of Primatology*, *73*(3), 291–303. doi:10.1002/ajp.20896

Jungers, W. L., Harcourt-Smith, W. E. H., Wunderlich, R. E., Tocheri, M. W., Larson, S. G., Sutikna, T., … Morwood, M. J. (2009). The foot of Homo floresiensis. *Nature*, *459*(7243), 81–4. doi:10.1038/nature07989

Kaifu, Y., Baba, H., Sutikna, T., Morwood, M. J., Kubo, D., Saptomo, E. W., … Djubiantono. (2011). Craniofacial morphology of Homo floresiensis: description, taxonomic affinities, and evolutionary implication. *Journal of Human Evolution*, *61*(6), 644–82. doi:10.1016/j.jhevol.2011.08.008

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., & Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, *32*(Database issue), D493–D496. doi:10.1093/nar/gkh103

Keinan, A., Mullikin, J. C., Patterson, N., & Reich, D. (2007). Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet*, *39*(10), 1251–1255. Retrieved from http://dx.doi.org/10.1038/ng2116

Kuhlwilm, M., Gronau, I., Hubisz, M. J., de Filippo, C., Prado-Martinez, J., Kircher, M., … Castellano, S. (2016). Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature*, *530*(7591), 429–433. doi:10.1038/nature16544

Lachance, J., Vernot, B., Elbers, C. C., Ferwerda, B., Froment, A., Bodo, J. M., … Tishkoff, S. a. (2012). Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell*, *150*(3), 457–469. doi:10.1016/j.cell.2012.07.009

Laplante, M., and Sabatini, D.M. (2012). mTOR signaling in growth control and disease. Cell 149,
274–293. Wullschleger, S., Loewith, R., and Hall, M.N. (2006). TOR signaling in growth and
metabolism. Cell 124, 471–484.

Larick, R., & Ciochon, R. L. (2015). Early hominin biogeography in Island Southeast Asia. *Evolutionary Anthropology*, *24*(5), 185–213. doi:10.1002/evan.21460

Larick, R., Ciochon, R. L., Zaim, Y., Sudijono, Suminto, Rizal, Y., … Heizler, M. (2001). Early Pleistocene 40Ar/39Ar ages for Bapang Formation hominins, Central Jawa, Indonesia. *Proceedings of the National Academy of Sciences* , *98* (9), 4866–4871. doi:10.1073/pnas.081077298

Larson, S. G., Jungers, W. L., Morwood, M. J., Sutikna, T., Jatmiko, Saptomo, E. W., … Djubiantono, T. (2007). Homo floresiensis and the evolution of the hominin shoulder. *Journal of Human Evolution*, *53*(6), 718–731. doi:http://dx.doi.org/10.1016/j.jhevol.2007.06.003

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., … Krause, J. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, *513*(7518), 409–413. doi:10.1038/nature13673

Leavesley, M. G. , Chappell J. (2004).Buang Merabak: Additional early radiocarbon evidence of the colonisation of the Bismarck Archipelago, Papua New Guinea. *Antiquity* 78, 301 available at http://www.antiquity.ac.uk/projgall/leavesley/

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. doi:10.1093/bioinformatics/btp324

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–6. doi:10.1038/nature10231

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Subgroup, 1000 Genome Project Data Processing. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* , *25* (16), 2078–2079. doi:10.1093/bioinformatics/btp352

Lipson, M., Loh, P.-R., Patterson, N., Moorjani, P., Ko, Y.-C., Stoneking, M., … Reich, D. (2014). Reconstructing Austronesian population history in Island Southeast Asia. *Nat Commun*, *5*. Retrieved from http://dx.doi.org/10.1038/ncomms5689

Liu, H., Prugnolle, F., & Manica, A. (2006). A Geographically Explicit Genetic Model of Worldwide, *79*(August), 230–237.

Liu, X., & Fu, Y.-X. (2015). Exploring population size changes using SNP frequency spectra. *Nature Genetics*, *47*(5), 555–559. doi:10.1038/ng.3254

Long, X., Spycher, C., Han, Z. S., Rose, A. M., Müller, F., & Avruch, J. (2002). TOR Deficiency in C. elegans Causes Developmental Arrest and Intestinal Atrophy by

Inhibition of mRNA Translation. *Current Biology*, *12*(17), 1448–1461. doi:http://dx.doi.org/10.1016/S0960-9822(02)01091-6

Lordkipanidze D, Ponce de León MS, Margvelashvili A, Rak Y, Rightmire GP, Vekua A, Zollikofer CP (2013) A complete skull from Dmanisi, Georgia, and the evolutionary biology of early Homo. Science. 342:326-31

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics (Oxford, England)*, *26*(22), 2867–73. doi:10.1093/bioinformatics/btq559

Margvelashvili A. et al. (2013) Tooth wear and dentoalveolar remodeling are key factors of
morphological variation in the Dmanisi mandibles. *PNAS* 110:17278-83.

Martin, S. H., Davey, J. W., & Jiggins, C. D. (2014). Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Molecular Biology and Evolution* . doi:10.1093/molbev/msu269

McDougall, I., Brown, F. H., & Fleagle, J. G. (2005). Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature*, *433*(7027), 733–736. Retrieved from http://dx.doi.org/10.1038/nature03258

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. a. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–303. doi:10.1101/gr.107524.110

McQuillan, R., Leutenegger, A. L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L., … Wilson, J. F. (2008). Runs of Homozygosity in European Populations. *American Journal of Human Genetics*, *83*(3), 359–372. doi:10.1016/j.ajhg.2008.08.007

Mendez, F. L., Watkins, J. C., & Hammer, M. F. (2012). A haplotype at STAT2 introgressed from neanderthals and serves as a candidate of positive selection in Papua New Guinea. *American Journal of Human Genetics*, *91*(2), 265–274. doi:10.1016/j.ajhg.2012.06.015

Meyer, M., Kircher, M., Gansauge, M., Li, H., Racimo, F., Siebauer, M., … Reich, D. (2012). A High-Coverage Genome Sequence from an Archaic Denisovan Individual A High-Coverage Genome Sequence from an Archaic Denisovan Individual A High-Coverage Genome Sequence from an Archaic Denisovan Individual, 1–14. doi:10.1126/science.1224344

Morwood, M. J., Brown, P., Jatmiko, Sutikna, T., Saptomo, E. W., Westaway, K. E., … Djubiantono, T. (2005). Further evidence for small-bodied hominins from the Late Pleistocene of Flores, Indonesia. *Nature*, *437*(7061), 1012–7. doi:10.1038/nature04022

Morwood, M. J., Soejono, R. P., Roberts, R. G., Sutikna, T., Turney, C. S. M., Westaway, K. E., … Fifield, L. K. (2004). Archaeology and age of a new hominin from Flores in eastern Indonesia. *Nature*, *431*(7012), 1087–1091. Retrieved from http://dx.doi.org/10.1038/nature02956

Murray-Wallace CV, Woodroffe CD. (2014).Quaternary sea-level changes: a global perspective.Cambridge: Cambridge University Press.

O'Connor S. New evidence from East Timor contributes to our understanding of earliest modern human colonisation east of the Sunda Shelf. Antiquity. 2007;81:523–35.

O'Connor, S. (2007). New evidence from East Timor contributes to our understanding of earliest modern human colonisation east of the Sunda Shelf. *Antiquity*, *81*(December 2006), 523–535. doi:10.1017/S0003598X00095569

O'Connor, S. (2015) in Emergence and Diversity of Modern Human Behaviour in Paleolithic Asia (eds Yousuke Kaifu et al.) 214–224. Texas A&M Univ. Press.

Oldham, S., Montagne, J., Radimerski, T., Thomas, G., & Hafen, E. (2000). Genetic and biochemical characterization of dTOR, the Drosophila homolog of the target of rapamycin. *Genes & Development* , *14* (21), 2689–2694. doi:10.1101/gad.845700

Oxnard, C., Obendorf, P.J., Kefford, B.J., 2010. Post-cranial skeletons of hypothyroid cretins show
a similar anatomical mosaic as Homo floresiensis. PLoS One 5, e13018.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., … Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*(3), 1065–1093. doi:10.1534/genetics.112.145037

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), 2074–2093. doi:10.1371/journal.pgen.0020190

Perry, G. H., & Dominy, N. J. (2009). Evolution of the human pygmy phenotype. *Trends in Ecology and Evolution*, *24*(4), 218–225. doi:10.1016/j.tree.2008.11.008

Perry, G. H., Foll, M., Grenier, J.-C., Patin, E., Nédélec, Y., Pacis, A., … Barreiro, L. B. (2014). Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. *Proceedings of the National Academy of Sciences* , *111* (35), E3596–E3603. doi:10.1073/pnas.1402875111

Peter, B. M. (2016). Admixture, Population Structure and F-Statistics. *Genetics*. Retrieved from http://www.genetics.org/content/early/2016/02/03/genetics.115.183913.abstract

Plagnol, V., & Wall, J. D. (2006). Possible ancestral structure in human populations. *PLoS Genetics*, *2*(7). doi:10.1371/journal.pgen.0020105

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., … Eichler, E. E. (2014). *The complete genome sequence of a Neandertal from the Altai Mountains*. doi:10.1038/nature12886.The

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., … Ostell, J. M. (2014). RefSeq: an update on mammalian reference sequences. *Nucleic Acids Research*, *42*(Database issue), D756–D763. doi:10.1093/nar/gkt1114

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. a R., Bender, D., … Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–75. doi:10.1086/519795

Pusey A., Wolf M., (1996). Inbreeding avoidance in animals. Trends Ecol. Evol. 11: 201–206

Qin, P., & Stoneking, M. (2015). Denisovan ancestry in East Eurasian and Native American populations. *Molecular Biology and Evolution*, *32*(10), 2665–2674. doi:10.1093/molbev/msv141

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* , *26* (6), 841–842. doi:10.1093/bioinformatics/btq033

Racimo, F., Sankararaman, S., Nielsen, R., & Huerta-Sanchez, E. (2015). Evidence for archaic adaptive introgression in humans. *Nat Rev Genet*, *16*(6), 359–371. Retrieved from http://dx.doi.org/10.1038/nrg3936

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. a, Feldman, M. W., & Cavalli-Sforza, L. L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(44), 15942–15947. doi:10.1073/pnas.0507611102

Rasmussen, M. D., Hubisz, M. J., Gronau, I., & Siepel, A. (2014). Genome-Wide Inference of Ancestral Recombination Graphs. *PLoS Genet*, *10*(5), e1004342. Retrieved from http://dx.doi.org/10.1371%2Fjournal.pgen.1004342

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., … Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature*, *411*(6834), 199–204. Retrieved from http://dx.doi.org/10.1038/35075590

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., … Pääbo, S. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, *468*(7327), 1053–1060. doi:10.1038/nature09710

Reich, D., Patterson, N., Kircher, M., Delfin, F., Nandineni, M. R., Pugach, I., … Stoneking, M. (2011). Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *American Journal of Human Genetics*, *89*(4), 516–528. doi:10.1016/j.ajhg.2011.09.005

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. (2009). Reconstructing Indian population history. *Nature*, *461*. doi:10.1038/nature08365

Roberts, R. G., Jones, R., & Smith, M. A. (1990). Thermoluminescence dating of a 50,000-year-old human occupation site in northern Australia. *Nature*, *345*(6271), 153–156. Retrieved from http://dx.doi.org/10.1038/345153a0

Rodriguez-Flores, J., Fakhro, K., Agosto-Perez, F., Ramstetter, M., Arbiza, L., Vincent, T., … Mezey, J. (2016). Indigenous Arabs are descendants of the earliest split from ancient Eurasian populations. *Genome Research*, 1–12. doi:10.1101/gr.191478.115

Rosenberg, N. A. (2004). Distruct: a program for the graphical display of population. *Mol Ecol Notes*, *4*. doi:10.1046/j.1471-8286.2003.00566.x

Ross, M. (2005) Pronouns as a preliminary diagnostic for grouping Papuan languages in: Papuan pasts: cultural, linguistic and biological histories of Papuan-speaking peoples. A. Pawley, R. Attenborough, J. Golson, R. Hide, editors. (Pacific Linguistics ,Canberra), pp. 15–65.

Ross, M, Pawley, A. Osmond, M. (2007) The Lexicon of Proto Oceanic: The culture and environment of ancestral Oceanic society: 2 The physical environment (ANU Press; http://www.jstor.org/stable/j.ctt24hfkc).

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., … Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, *507*(7492), 354–7. doi:10.1038/nature12961

Scally, A., & Durbin, R. (2012). Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews. Genetics*, *13*(10), 745–53. doi:10.1038/nrg3295

Seboun, E., Barbaux, S., Bourgeron, T., Nishi, S., Algonik, A., Egashira, M., … Kasahara, M. (1997). Gene Sequence, Localization, and Evolutionary Conservation of DAZLA,a Candidate Male Sterility Gene. *Genomics*, *41*(2), 227–235. doi:http://dx.doi.org/10.1006/geno.1997.4635

Sémah, F., Saleki, H., Falguères, C., Féraud, G., & Djubiantono, T. (2000). Did Early Man reach Java during the Late Pliocene? *Journal of Archaeological Science*, *27*(9), 763–769. doi:10.1006/jasc.1999.0482

Serre, D., Langaney, A., Chech, M., Teschler-Nicola, M., Paunovic, M., Mennecier, P., … Paabo, S. (2004). No Evidence of Neandertal mtDNA Contribution to Early Modern Humans. *PLoS Biol*, *2*(3), e57. Retrieved from http://dx.doi.org/10.1371%2Fjournal.pbio.0020057

Sheehan, S., Harris, K., & Song, Y. S. (2013). Estimating Variable Effective Population Sizes from Multiple Genomes : A Sequentially Markov, *194*(July), 647–662. doi:10.1534/genetics.112.149096

Skoglund, P., & Jakobsson, M. (2011). Archaic human ancestry in East Asia. *Proceedings of the National Academy of Sciences*, *108*(45), 18301–18306. doi:10.1073/pnas.1108181108

Skoglund, P., Mallick, S., Bortolini, M. C., Chennagiri, N., Hünemeier, T., Petzl-Erler, M. L., … Reich, D. (2015). Genetic evidence for two founding populations of the Americas. *Nature*, *525*(7567), 104–110. doi:10.1038/nature14895

Spoor, F., Leakey, M. G., Gathogo, P. N., Brown, F. H., Anton, S. C., McDougall, I., … Leakey, L. N. (2007). Implications of new early Homo fossils from Ileret, east of Lake Turkana, Kenya. *Nature*, *448*(7154), 688–691.

Spriggs M. 2000. Out of Asia? The spread of Southeast Asian Pleistocene and Neolithic maritime cultures in Island Southeast Asia and the western Pacific. In: O'Connor S, Veth P, editors. East of Wallace's line: studies of past and present maritime cultures of the Indo-Pacific region. Rotterdam (the Netherlands): A.A. Balkema. p. 51–75

Stewart, J. R., & Stringer, C. B. (2012). Human evolution out of Africa: the role of refugia and climate change. *Science (New York, N.Y.)*, *335*(6074), 1317–21. doi:10.1126/science.1215627

Summerhayes, G. R. Island Melanesian Pasts — A view from archaeology. in: Genes, language, and culture history in the Southwest Pacific. J.S. Friedlaender, editor. (Oxford University Press, New York, NY, 2007; http://catalogue.nla.gov.au/Record/3992849), pp. 10-35.

Swisher, C. C., Curtis, G. H., Jacob, T., Getty, a G., Suprijo, a, & Widiasmoro. (1994). Age of the earliest known hominids in Java, Indonesia. *Science (New York, N.Y.)*. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8108729

Swisher, C. C., Rink, W. J., Antón, S. C., Schwarcz, H. P., Curtis, G. H., Suprijo, a, & Widiasmoro. (1996). Latest Homo erectus of Java: potential contemporaneity with Homo sapiens in southeast Asia. *Science (New York, N.Y.)*, *274*(5294), 1870–1874. doi:10.1126/science.274.5294.1870

Tassi F., Ghirotto S., Mezzavilla M., Torres Vilaça S., De Santi L., Barbujani G. (2015) Early modern human dispersal from Africa: Genomic evidence for multiple waves of migration. *Investigative Genetics* 6:13

Trinkaus, E., Moldovan, O., Milota, ştefan, Bîlgăr, A., Sarcina, L., Athreya, S., ⋯ van der Plicht, J. (2003). An early modern human from the Peştera cu Oase, Romania. *Proceedings of the National Academy of Sciences* , *100* (20), 11231–11236. doi:10.1073/pnas.2035108100

van den Bergh, G. D., Li, B., Brumm, A., Grün, R., Yurnaldi, D., Moore, M. W., … Morwood, M. J. (2016). Earliest hominin occupation of Sulawesi, Indonesia. *Nature*, *529*(7585), 208–11. doi:10.1038/nature16448

Vekua, a. (2002). A New Skull of Early Homo from Dmanisi, Georgia. *Science*, *297*(5578), 85–89. doi:10.1126/science.1072953

Vernot, B., & Akey, J. M. (2014). Resurrecting surviving Neandertal lineages from modern human genomes. *Science*, *343*. doi:10.1126/science.1245938

Walker, A.and R. Leakey (1993). The Nariokotome Homo erectus skeleton. Cambridge, Harvard Univ. Press.

Wall, J. D. (2000). Detecting ancient admixture in humans using sequence polymorphism data. *Genetics*, *154*(3), 1271–9.

Wall, J. D., Yang, M. a., Jay, F., Kim, S. K., Durand, E. Y., Stevison, L. S., … Slatkin, M. (2013). Higher levels of Neanderthal ancestry in east Asians than in Europeans. *Genetics*, *194*(1), 199–209. doi:10.1534/genetics.112.148213

Wang, S., Lachance, J., Tishkoff, S. a., Hey, J., & Xing, J. (2013). Apparent variation in Neanderthal admixture among African populations is consistent with gene flow from non-African populations. *Genome Biology and Evolution*, *5*(11), 2075–2081. doi:10.1093/gbe/evt160

Weber J, Czarnetzki A, Pusch CM. 2005. Comment on ''The brain of LB1. Homo floresiensis.'' Science 310:236.

Weston, E. M., & Lister, a M. (2009). Insular dwarfism in hippos and a model for brain size reduction in Homo floresiensis. *Nature*, *459*(7243), 85–88. doi:10.1038/nature07922

Wickler, S., Spriggs, M. (1988). Pleistocene human occupation of the Solomon Islands, Melanesia. *Antiquity* 62, 703–706 doi:10.1017/S0003598X00075104

Wright S., (1922). Coefficients of inbreeding and relationships. Am. Nat. 56: 330–339

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(1), 3–36. doi:10.1111/j.1467-9868.2010.00749.x

Wurm, S. A. (1975) Papuan languages and the New Guinea linguistic scene (Dept. of Linguistics, School of Pacific Studies, Australian National University, Canberra, http://nla.gov.au/nla.cat-vn2122276).

Xu, S., Pugach, I., Stoneking, M., Kayser, M., & Jin, L. (2012). Genetic dating indicates that the Asian-Papuan admixture through Eastern Indonesia corresponds to the Austronesian expansion. *Proceedings of the National Academy of Sciences*, *109*(12), 4574–4579. doi:10.1073/pnas.1118892109

Yang, X., & Xu, S. (2011). Identification of close relatives in the HUGO Pan-Asian SNP database. *PloS One*, *6*(12), e29502. doi:10.1371/journal.pone.0029502

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*(24), 3326–3328. doi:10.1093/bioinformatics/bts606

Zinner, D., Arnold, M. L., & Roos, C. (2011). The strange blood: natural hybridization in primates. *Evolutionary Anthropology*, *20*(3), 96–103. doi:10.1002/evan.20301

1000 Genomes Project Consortium, et al, (2015) A global reference for human genetic variation. *Nature.* 526, 68-74