New Insights into Human Migration, Demography and Adaptation of Indian and South Asian populations from genome analyses

Mayukh Mondal

TESI DOCTORAL UPF / 2016
DIRECTOR DE LA TESI
Prof. Jaume Bertranpetit and Dr. Ferran Casals
DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

Universitat Pompeu Fabra
Barcelona

To my recently deceased grandfather who was always earnest about my education. I could not attend his funeral due to this thesis and visa problem. But knowing him, I know he would understand my situation and forgive me.

“You should not work on weekends. You should enjoy the life also. If you work too much, you will go crazy.”

-Ferran Casals on the very first day I landed in Barcelona after a long flight

“Mayukh, if you are right we’ll share the glory. If you are wrong, I will take the blame and protect you from others”

-Jaume Betranpetit at the time of big debate for our main paper

“Sometime a single sentence can define a person”

-Mayukh Mondal on a neverending fight with my “Schrödinger's” girlfriend

Acknowledgements

This is the beginning of the "end" product from a wonderful Doctor of Philosophy (PhD) thesis project that I was lucky to be involved for last 4 years (2012-2016) in IBE under the supervision of Prof. Jaume Bertranpetit and Dr. Ferran Casals. Rather than raining thanks in every direction, I would stay true to the definition of Acknowledgements (i.e. recognition of the importance or quality of something) and write what I believe should be noted as important to shape this PhD thesis work.

I guess this project would not be what it is today without the help of both of my supervisor Ferran and Jaume. Ferran was really involved in starting of this project and I discussed about every single graph I produced (and on an average I would produce a new graph every day so I have discussed with him every day). He was always accommodating and we discussed a lot. Although later he got more involved with "Genomic Core Facility" but at that time I have already learnt a lot and could survive without his help in every single step. In later years, Ferran's absence was replaced by Jaume's presence and we have load of discussions very often. I really appreciate Jaume's intuition about what is right or wrong. Whenever I presented some new idea he was the first person to understand what I meant to say and helped me a lot how to make it simple so that everybody can understand. He never forces his idea on me rather always encouraged me to pursue the idea in more depth. My guess without these encouragements I could not have done so much interesting yet controversial works. I would also like to thank my ex (master's) supervisor Prof. Partha P Majumder. He was the first person to recommend me to do this Indo-Spanish project. Sometime I get confused if I am from the Spanish side or from the Indian side in this Indo-Spanish collaboration. I guess the most important lesson I learnt from these three people is that they stood by me at the time of need rather than putting all the blame on me. A true leader not only shares glory but also failures.

I would like to thank other members of our group as well; Giovanni to help me in the starting of the project and to do the complicated variant calling pipeline, Marc and Pierre helping me in the selection pipeline and Hafid for the knowledge of biostatistics. It was a real pleasure to meet our new young lab members: Begona, Sandra, Jessica, Pablo and Apostolia. Generally, I am a lone wolf doing my project alone but this is the end of my PhD and I am going to do some Postdoc and eventually become Principal Investigator (if have the chance). Thus, I need to improve my skills on distributing knowledge and leadership. I learnt a lot how to make simple yet clear conceptual ideas and distribute to younger generation when teaching them new concept or tool. In that sense, they were my first students. I learnt an important lesson that doing something for you and making other do that is not same, which I know, will help me in my future endeavours.

I would also like to thank Txema to tolerate me in the earlier days of cluster computing in IBE. Very often, I put the cluster to its knees and get some "interesting" mails from him. By reading these emails it is difficult assess if he was angry or just joking (not so much different when you talk with him personally), nonetheless it is interesting to read. I am sure I will never forget him because of those emails. Also would like to thank Juanma who is handling the cluster now and running it very smoothly. It was a wonderful experience to be part of IBE with some good friends (especially Raj, Nino, Juan, Guillem, Marco etc) to show me the

nightlife of Barcelona so that I never felt lonely here.
Last but not the least I should acknowledge Barcelona as a whole. It is a wonderful city, situated just beside the beach (which I really enjoyed and let me keep my tan) and its wonderful weather. It is a big cultural shock for an Indian who lived all his life with in India. India by no means has homogenous culture but when I came out of India and stayed here, it changed a lot of my perspective about the world. I guess people do not realize the biases placed by their own society because they have never left it. To truly know the ingrained biases placed by our own society, people should leave and live in a completely different society for a substantial amount of time. It would make them humbler about the diversity and enlighten them to a new and unbiased perspective of life, which is not possible by living within their own culture.

And the real acknowledgement goes to:

This PhD was supported by "FI-Agaur" PhD fellowship and the research was supported by the joint Spain–India bilateral grant PRI-PIBIN-2011-0942 from the Ministerio de Economía y Competitividad (Spain).

Abstract

Human genome project published their first human whole genome sequence on 2001 at the cost of billions of dollars. Since, the cost of sequencing is decreasing faster than Moore's law. Now, we not only have sequenced thousands of modern humans' whole genome, we also obtained whole genome sequences of extinct hominin and other ancient modern humans with relatively good quality. These sequences granted us some unexpected results: like how recently modern humans left Africa and populated around all over the world (which is called recent African origin model) while doing so how they have admixed with multiple hominin populations. Until now modern biology (unlike physics) always dominated by empirical results compared to theoretical concepts, which forces people to perceive biology as a descriptive science. As we are obtaining more and more data every day, it is now time to push our theoretical concepts before empirical results in biology. Here in this thesis, we provided deeper knowledge about ancestry of Indian, Asian and Pacific populations. We were also able to reveal an unknown hominin population existed even before it is sequenced. In addition to these, we demonstrated strong natural selection could change human morphology drastically in a short period.

Resum

El projecte del genoma humà va publicar la primera seqüència completa del genoma humà el 2001 amb un cost de milers de millions de dòlars. Després d'això, el cost de la seqüenciació està disminuint més ràpid que la llei de Moore. Actualment no només tenim la seqüència de del genoma humà, sinó que tenim la de molts humans i d'homínids extingits amb una qualitat relativement bona. L'estudi de les seqüències de molts genomes humans varen proporcionar la base per postular que els humans moderns es varen originar a Àfrica, i en la sortida d'Àfrica (Out Of Africa) varen poblar la resta del món, amb una certa barreja amb diferents poblacions d'homínids. La base del treball en biologia i en genòmica evolutiva ha estat fonamentalment empírica (a diferència de la física), però actualment la disponibilitat de moltes dades permet empenyer la recerca cap a aspectes molt més analítics: aquest és l'enfocament del nostre treball en seqüències de DNA. Aquí, en aquesta tesi, hem proporcionat un coneixement més profund sobre l'origen i l'ascendència de poblacions indígenes, d'Àsia i del Pacífic, centrant-nos en la India continental i especialment en les Illes Andaman. També hem estat capaços de revelar l'existència d'una població d'homínids desconeguts que es va barrejar amb els ancestres d'aquestes poblacions. A més, hem demostrat que una forta selecció natural pot canviar dràsticament la morfologia humana en un curt període de temps i que explicaria la morfologia pigmea del pobladors de les illes Andaman.

সারাংশ

হিউম্যান জিনোম প্রজেক্ট চল্লিশ কোটি ডলার খরচে করে সম্পূর্ণ মানব জিনোম সিকোয়েন্স প্রথম প্রকাশিত করে ২০০১ সালে। এরপর, সিকোয়েন্সিংর খরচ মুরের আইন তুলনাতেও দ্রুততর হারে কমেছে। হাজার হাজার বর্তমান মানুষের পুরো জিনোম সিকোয়েন্স ছাড়াও আমরা অপেক্ষাকৃত ভাল গুণমান যুক্ত বিলুপ্ত হোমিনিন এবং প্রাচীন বিলুপ্ত মানুষের পুরো জিনোম সিকোয়েন্সও প্রাপ্ত করেছি। এই সিকোয়েন্সগুলি থেকে উঠে এসেছে অপ্রত্যাশিত কিছু তত্ত্ব: যেমন সম্প্রতি কিভাবে আধুনিক মানুষেরা আফ্রিকা ছেড়ে বেরিয়ে এসে সাড়া বিশ্বে ছড়িয়ে পড়েছে (যাকে সাম্প্রতিক আফ্রিকান বংশোদ্ভূত মডেল বলা হয়) এবং এটা করতে গিয়ে কিভাবে তারা একাধিক হোমিনিন জনগোষ্ঠীর সঙ্গে সংমিশ্রিত হয়েছে। কিছুদিন আগেও আধুনিক জীববিজ্ঞান (পদার্থবিদ্যার মতোন নয়) বেশির ভাগ সময় তাত্ত্বিক ধারণার থেকে পরীক্ষালব্ধ ফলাফলের উপরই বেশি জোর দিতো, যা কিনা জীববিদ্যাকে একটি বর্ণনামূলক বিজ্ঞানের রূপ দিতো সাধারন জনগনের কাছে। যেহেতু আমরা প্রতিদিন বিপুল মাত্রায় তথ্য প্রাপ্তি করছি, জীববিজ্ঞানে পরীক্ষালব্ধ ফলাফলের আগে তাত্ত্বিক ধারণাকে এগিয়ে রাখার সময় এসে গেছে। এখানে এই গবেষণামূলক প্রবন্ধে আমরা ভারতীয়, এশীয় ও প্রশান্ত মহাসাগরীয় জনগণের পূর্বপুরুষদের সম্পর্কে গভীর জ্ঞান প্রদান করেছি। আমরা একটি অজানা হোমিনিডের অস্তিত্ব প্রকাশ করতে পেরেছি সিকোয়েন্সের ও পূর্বে। এছাড়াও আমাদের গবেষণা প্রদর্শন করেছে কিভাবে শক্তিশালী প্রাকৃতিক নির্বাচন স্বল্প সময়ের মধ্যে মানুষের আয়তন পরিবর্তন করতে পারে।

Preface

Research on ancestry of European populations drastically increased in last 4 or 5 years via whole genome sequencing (mainly done by ancient genomes). These genomes changed or challenged some old ideas about how modern humans started to populate Europe. Although these studies were interesting, unfortunately deep research on other populations (African, Asian, Pacific etc.) lacked behind. These non-European populations are generally used as a backbone for European ancestry until recently.

Here in this thesis we mainly concentrated on Indian populations, but secondarily also on Aboriginal Australians and Pacific populations. Indian populations are fascinating. With more than 1 billion of Individuals (⅙ th of whole world populations) and a complex ancestry, they remained underrepresented in population history studies. We tried to delimit ancestry of Indian populations first with the help of 120 Genotype and Exome data from main two Indian populations (North and South) and later using 70 whole genome sequences from diverse geographical regions, linguistic affiliations and social categories. We also attempted to look for how recent adaptation shaped these populations. In addition to that, we also outlined Y-chromosome ancestry of these populations. Unexpectedly we discovered an unknown hominin population introgressed in Andamanese, which was later shown to be introgressed in Asian and Pacific populations also. As detecting unknown hominin population is not well developed, we needed to develop our own method to detect that. Our simulation models revealed this method is good enough to detect unknown hominin populations and can be implemented on any population.

List of Publications

1. Juyal, Garima et al. "Population and genomic lessons from genetic analysis of two Indian populations." Human genetics 133.10 (2014): 1273-1287.
2. Mondal, Mayukh et al. "Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation." Nat Genet. 2016 Sep;48(9):1066-1170. doi: 10.1038/ng.3621. Epub 2016 Jul 25.
3. Mondal, Mayukh et al. "Further confirmation for unknown archaic ancestry in Andaman and South Asia." BioRxiv doi: http://dx.doi.org/10.1101/071175

In Preparations

1. East Asian introgression: an overall vision of ancient introgressions in the human lineage.
2. Ancestry and unknown Hominin introgression in Australian genomes.
3. Y chromosome profile of Indian continental populations and ancestry dilemma of Andamanese Populations.

Table of Content

1. Introduction

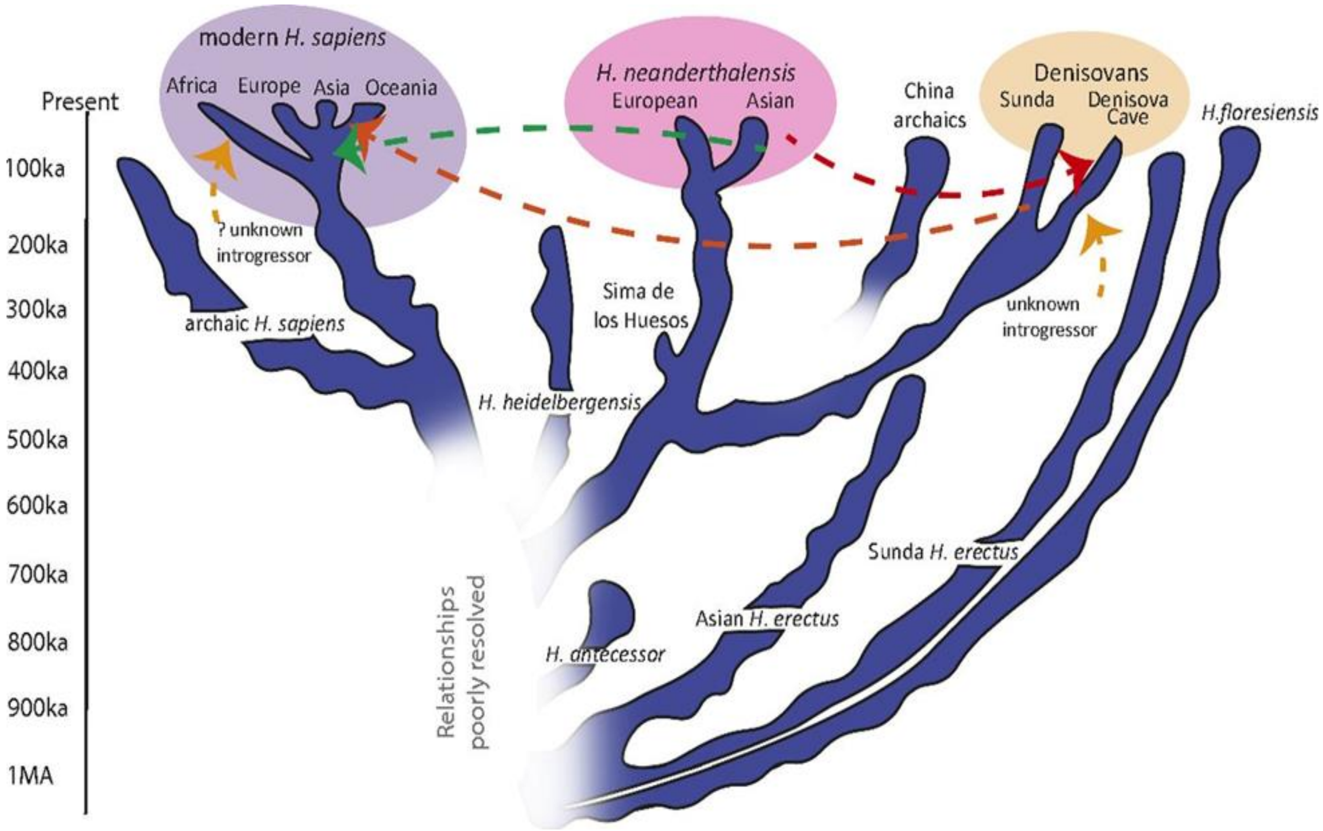
1.1 Origin of Humans as a species

The last living common ancestor of humans (*Homo genus*) is chimpanzees (*Pan genus*). It is estimated to be separated from us around less than 6 million years ago by genetic studies (Patterson, Richter, *et al.*, 2006; Prado-Martinez *et al.*, 2013), whereas fossils data showed a much older date for speciation around 7-10 million years ago (Suwa *et al.*, 2007; White *et al.*, 2009). This discrepancy between genetic studies and fossils dates was sometimes explained by differences in estimation of germline mutation rate (Moorjani *et al.*, 2016) or generation time estimation (Langergraber and Prüfer, 2012). No matter the true date of divergence, scientific community agrees that the closest living species of Homo sapiens is Chimpanzee and Bonobo. This is especially interesting as it can give us clues of being "human". How we developed the acute amount of intelligence - which made us one of the most dominant species on earth - whereas other ape species are in the danger zone of extinction (IUCN, 2015). It looks like the evolution of intelligence happened on earth only once (at least the level of intelligence where in every generation we improve upon the previous generation without changing our genome, a la cultural evolution). Although some other primates have known to have cultural transmission, in case of humans, the sheer amount of information transmission is unprecedented and it is increasing exponentially every year especially after the advent of digital age. Impact of this kind of intelligence is a blessing (e.g. increase in lifespan through use of antibiotics) or curse (e.g. increase global temperature and thus dooming all living species) is not a scientific topic for this thesis but as a scientist, I truly admire human intelligence, which also compelled me to do my PhD in human evolution.

1.2 Humans subgenus and modern humans

All humans around the world (around seven billion in total) are believed to be related to each other around 200 kilo years ago (kya) and came from Africa. This hypothesis is called "recent African origin model" and is supported by Mitochondrial Deoxyribonucleic acid (DNA) (Soares *et al.*, 2009), Y chromosome analysis (Poznik *et al.*, 2016), Autosomal analysis (Li and Durbin, 2011) as well as fossils data (McDougall *et al.*, 2005). Interestingly anthropologist discovered a lot of humanoid bones much older than that (>200 kya) in Eurasia, for a long time it was thought that our species developed independently around the world from this humanoid subspecies which are called archaic hominin (Wolpoff *et al.*, 2000). This hypothesis, which is called "multiregional origin of modern humans", was proven wrong after scientist could calculate most recent common ancestor from genetic data. All these analyses pointed that all living humans (at least the ones that are sequenced) have a most recent common ancestor around 200 kya. So all the living humans were named modern humans to distinguish them from other extinct hominin populations living outside Africa. To explain the hominin fossils found around the world before modern humans ever existed, the scientist proposed they came from our distant cousin, which left Africa much before modern humans (Figure 1.2.1). Although there is no consensus yet on where the common ancestor of all hominins originated, before sequencing of these hominin fossils, it was mainly thought that modern humans, when they came out of Africa (OOA) around ~70 kya, replaced already living hominin populations and force them to go extinct (Diamond, 2014). But after sequencing of these populations, we found that all OOA populations have different amount of introgression from different hominin populations [all OOA populations have introgression from Neanderthals and Pacific populations have introgression from Denisova (Green *et al.*, 2010; Meyer *et al.*, 2012)]. In this thesis, we argued that another hominin population existed which have introgressed in all Asian populations (Methods and Results Section 3.2-3.5). So although these populations [some of them even have bigger brains than us (de León *et al.*, 2008)] are now extinct, they still live inside us (i.e. inside the genome of modern humans) around the world.

Figure 1.2.1: A model of Humans evolution (Stringer et al. 2015).



1.3 Ancestry and Out of Africa dispersal of modern humans

All the modern humans are related to each other ~ 200 kya as discussed earlier. Around 70 kya (possibly due to climate change (Parton *et al.*, 2015)), modern humans started to disperse and few of them left Africa (Melé *et al.*, 2012; Li and Durbin, 2011). It is interesting to note that we have two competing hypotheses around OOA event for modern humans. One hypothesis is that only "one OOA" event had happened (~70 kya) which created all the diversity of OOA populations. The second hypothesis supports "two OOA" event for modern humans. The first OOA event happened earlier [>100 kya (Grün *et al.*, 2005; Rasmussen *et al.*, 2011; Kuhlwilm *et al.*, 2016)], followed a coastal route and populated South, South-East Asia and Pacific. The second OOA event happened later (~70 kya) which produced Europeans, East Asians and all the other populations except few isolated populations (Andamanese, Papuan, Australian etc.) and replaced the first OOA populations that were living there. In this thesis, we supported single OOA hypothesis (at least for the populations that were mentioned earlier to be created from first OOA event; i.e. Andamanese, Australian and Papuan). It is interesting to mention that remnant of first OOA modern humans were discovered inside Neanderthals (Kuhlwilm *et al.*, 2016) but not in modern humans by genetic analysis, suggesting humans are (were) extremely promiscuous or friendly to other populations. Nonetheless, it is agreed that most of the variation of OOA populations (Europeans, East Asians, Native American, Indians etc.) was created from a single OOA event happened ~ 70 kya. One explanation for this seemingly negligible footprint of first OOA populations might be that although the first OOA event happened, these individuals died out and thus were not able to leave any descendants in the living modern human populations. One of the causes for going completely extinct might be caused by the mount Toba volcanic eruption which occurred 75 kya (Ambrose, 1998).

Figure 1.3.1: One out of Africa Migration theory from Wikimedia (data source: Burenhult, Göran et al. 2000)

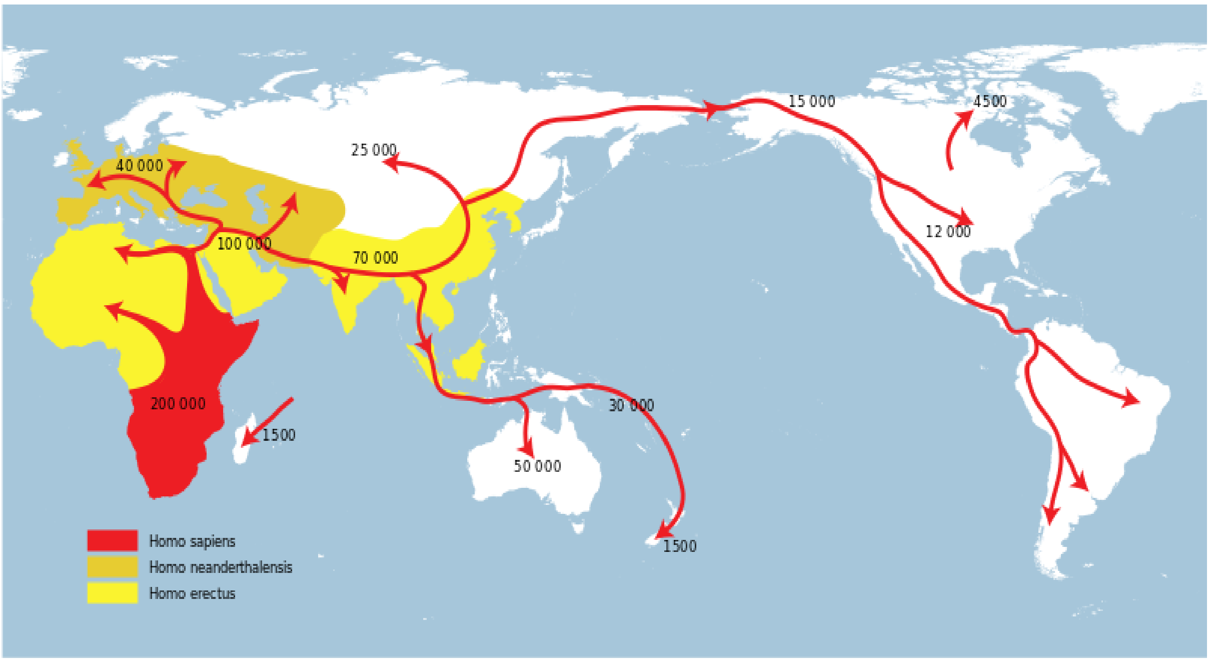
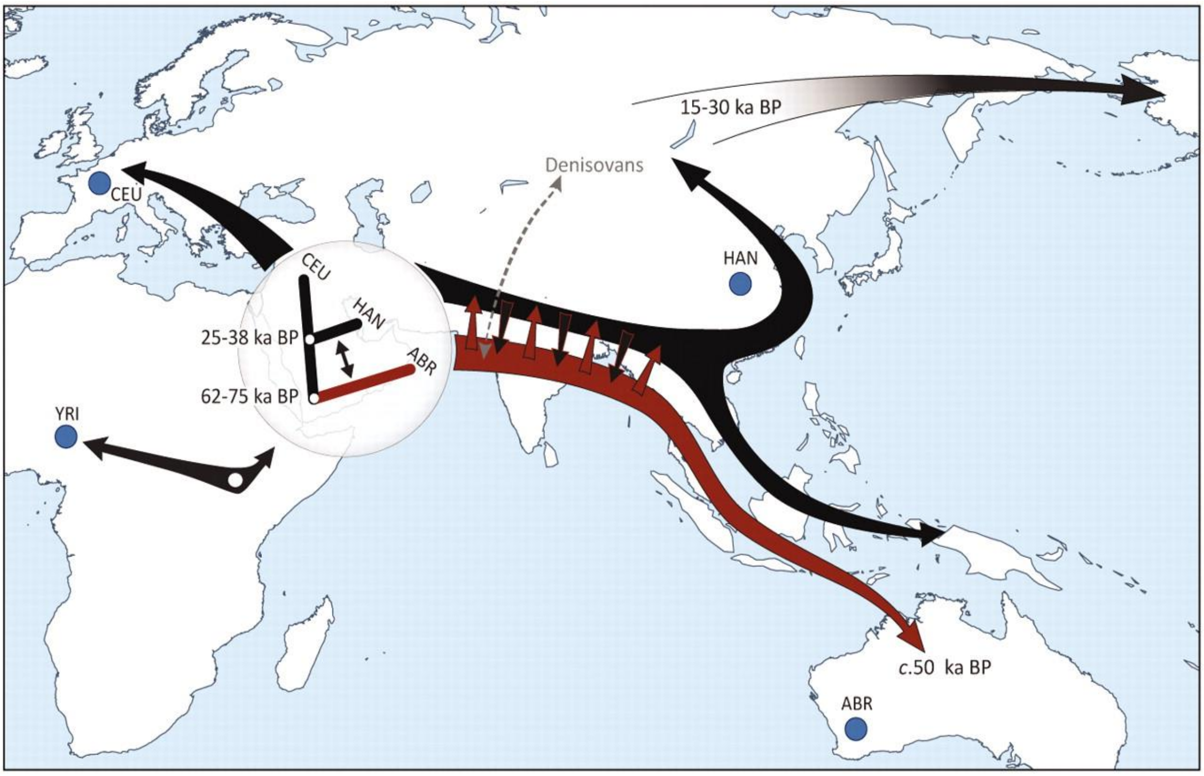


Figure 1.3.2: Two out of Africa Migration theory (Rasmussen, Morten et al. 2011)

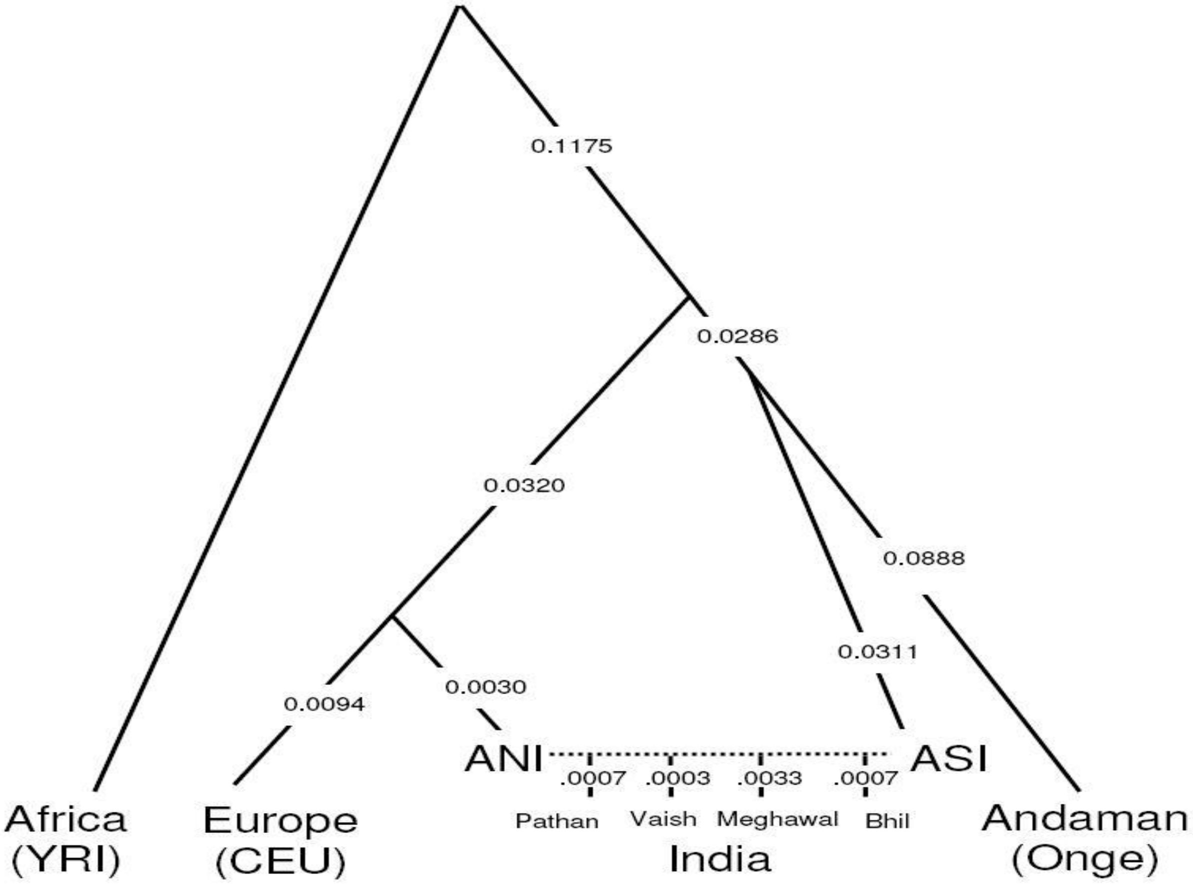


1.4 Indian populations

India is a big country with 1.21 billion people and 22 constitutionally accepted languages (with 30 languages spoken by more than 1 million people and 1635 “rationalise” languages in total (Government of India, 2011)). It has seventh biggest land mass as a country and seventh highest nominal Gross Domestic Product (GDP) in total (International Monetary Fund World Economic Outlook -April-2016). Although India has these big numbers to back it up, it is under-represented in population genetic studies until now. History of India is complicated. India has different language families with different ancestry because of multiple human migrations coming to India bringing those language families to India throughout time (from the historic and prehistoric era). Two most common language families are Indo-European and Dravidian, though other language families also exist in India with known and unknown ancestry (Tibeto-Burman, Austro-Asiatic, Andamanese languages etc.). In addition, some populations of India (i.e. Andamanese, Austro-Asiatic etc.) are extremely interesting due to their physical appearances and unique culture.

Indo-European languages, which are spoken by Indians, are attributed to be brought by Aryan migration around 2000 BCE thus related with other European languages (Bryant, 2003). Dravidian languages are thought to be present in India before that time and originated within India (Avari, 2007). Recent autosomal genetic studies from genotype data (Reich *et al.*, 2009; Basu *et al.*, 2015) reinforced correlation between the language spoken and genetic structure present in non-tribal populations of India. They have hypothesised that Indian populations have two component of ancestry. North India has a higher portion of Ancestral North Indian (ANI) component whereas South India has a higher amount of Ancestral South Indian (ASI) component. ANI can be correlated with Indo-European migration and ASI would be the Dravidian component, both of which admixing with varying degree throughout India (Figure 1.4.1). It is interesting to note that Basu et al. pointed out this component gradient is inadequate to describe the entire genetic component present in India. In addition to these two components, two other components also present in mainland India: Austroasiatic and Tibeto-Burman. They also noted that Andamanese are not genetically directly related with ASI component which was hypothesised previously (Reich *et al.*, 2009). We began with these discrepancies of Indian ancestry to delve deeper into it (Methods and Results Section 3.2). The ancestry of Austroasiatic and Tibeto-Burman will be addressed later.

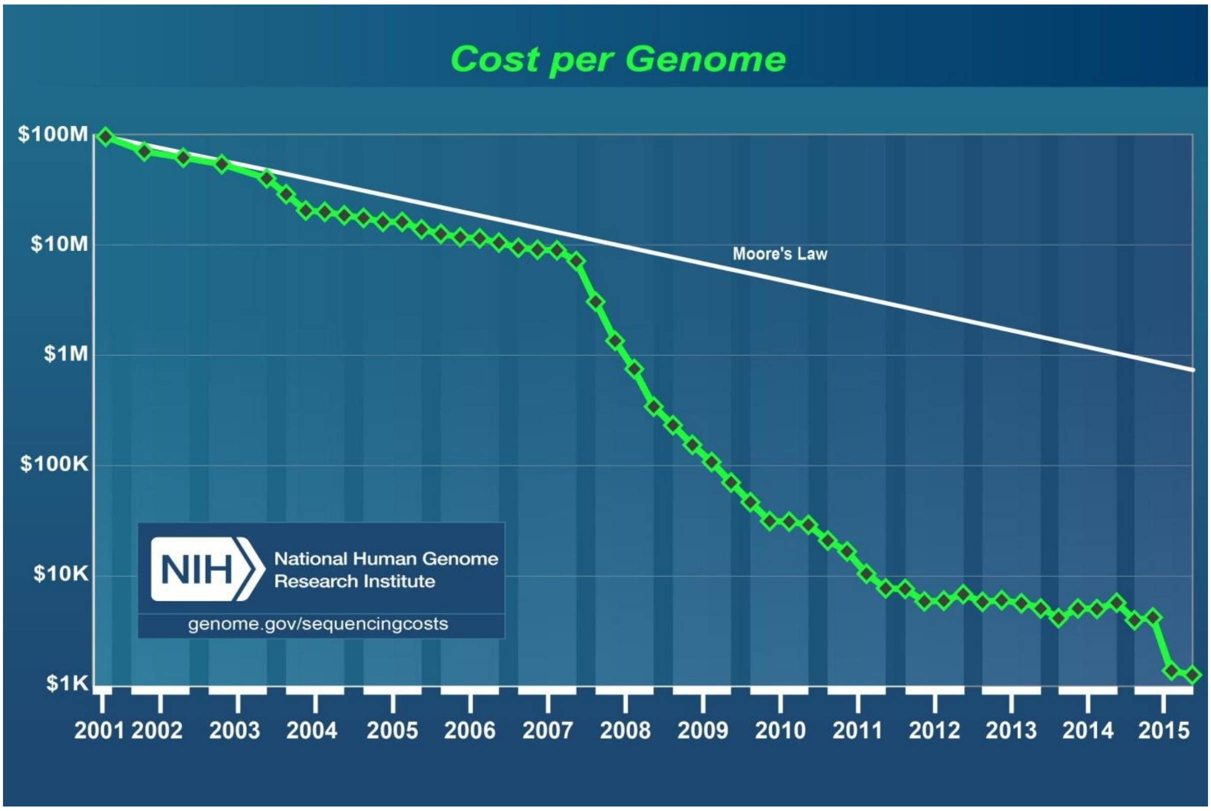
Figure 1.4.1: Model of 2 component of Indian Ancestry (Reich et al. 2009)



1.5 Explosion of genetic technologies

In 2001 The Human Genome Project Consortium published its draft genome after spending ~$4 billion dollars (Lander *et al.*, 2001). This was a big milestone for human population genetic studies, which changed completely after that. Just after publishing human reference genome, the cost of whole genome sequence plummeted faster than Moore's law (Figure 1.5.1). Nowadays we can sequence the whole genome of a human at a cost of $1000 and it is assumed that it will go lower in the future as sequencing technology become more efficient. The third generation sequencing technologies looks even more promising because of its low cost and long read sequencing (Schadt *et al.*, 2010).

Figure 1.5.1: Cost of sequence of whole genome human sequence (from NIH)



After the explosion of genome sequencing technology (due to massively parallel shotgun approach), it became apparent that sequencing individuals are relatively easy and cheap but storing, handling and understanding the big data (easily in Terabytes) is not anymore. With only a few individuals, it becomes close to terabytes of data with lots of data analysis processing involved in the middle. This amount of huge data analysis is beyond the power of desktop computers. Last few years we have multiple new tools, which can specifically handle this kind of data. With the availability of huge data sets, new tools are developed which can give an interesting insight into population's history.

1.6 Second Generation Sequencing Technology

Illumina became the reigning champion of second generation sequencing technology. The idea of Illumina is based on massively parallel sequencing technology using the sequencing by synthesis method. In a nutshell, DNA fragments are first attached to a surface and then cloned using Polymerase Chain Reaction (PCR) technology. After reaching enough density of clonal cluster for every DNA fragments, fluorescently tagged nucleotides are added. These fluorescently tagged nucleotides would shine a distinctive fluorescent light (four colours for four different type of nucleotides) when attaching to a DNA fragment. Cameras would detect this light. The attachment of this nucleotides cannot be random as they can only attach with the complementary nucleotides from the DNA fragment it is attaching, thus would produce a specific sequence of fluorescent light which then can be converted to DNA sequence read. The camera detection power goes down if the amount of light is coming too low. The previously mentioned PCR step was used just to improve the amount of DNA so that there would be more light to be emitted from every cluster for every cycle of DNA attachment.

1.7 Tools to delimit ancestry

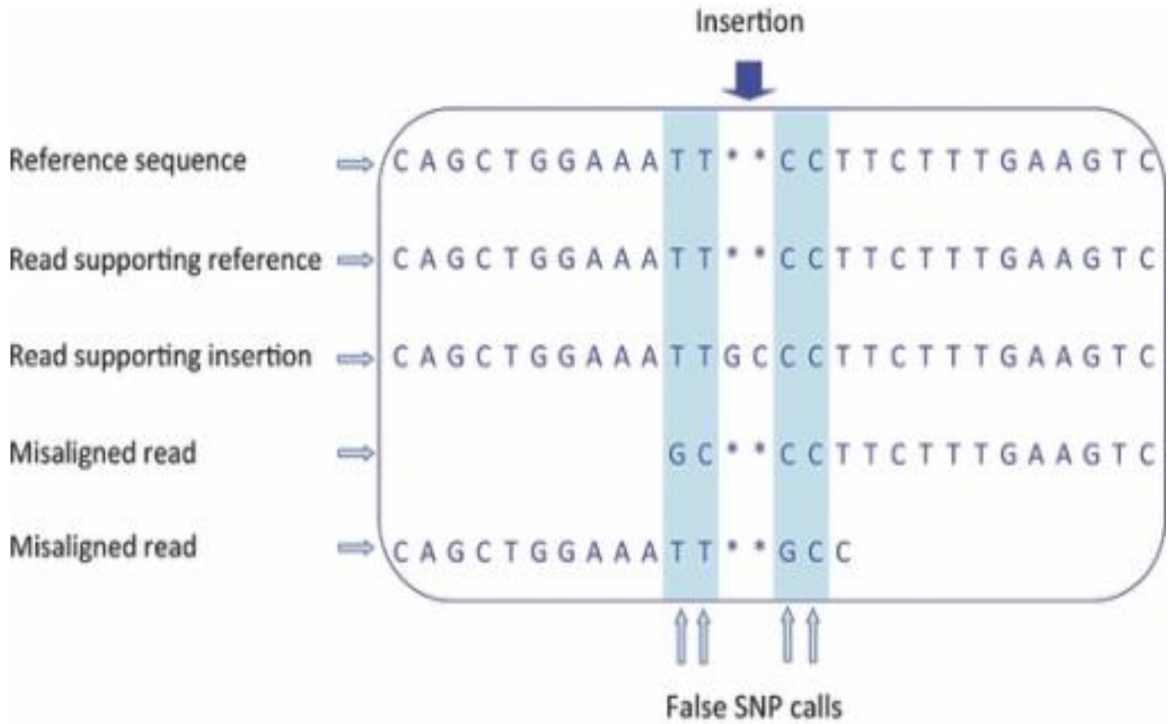
In this chapter, I will try to give basic ideas behind the methods that were used in this thesis.

1.7.1 Mapping Sequences (BWA) and Variant Calling

After generating billions of short sequence reads (in our case 90-100 nucleotides long pair end sequences) from Illumina, it is time to map them on known sequences which will eventually give an idea about what secret these sequences hold (a la shotgun method). We should think short sequences as words from a storybook. Individual words fail to make any sense. Only after putting them in certain order or context, the story can be understood. Likewise, we have to put these short sequences in proper order, which is done by mapping step. In 1990 the ground breaking paper for Basic Local Alignment Search Tool (BLAST) algorithm (Altschul *et al.*, 1990) was published, which improved bioinformatics analysis drastically. BLAST is more flexible and tries many combinations to map different sequences on each other. In the case of the human genome, the short sequences coming from Illumina is expected not to be very dissimilar from our reference as all humans on average have 99.9% sequence similarity (Jorde and Wooding, 2004). BLAST is good with mapping sequences with gaps, which is not necessary for human genome mapping because of high similarity. BLAST will waste a lot of time to search every possible combination including the repetitive sequences, whereas Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009) will only give the best possible result with less gap alignment and thus would reduce the time for mapping. This is a trade-off between accuracy and speed. Every Time we invoke a gap the combinations of mapping increases exponentially thus increase the timing also. In this case, BWA being strict saves lot of time. Thus, we use BWA instead of BLAST. BWA is much stricter to find similarity, but because of strictness, it is also faster to find a match for billions of sequences. BWA is less sensitive around highly diverged region (regions with more gaps), but doing a proper search for every short sequence would take time beyond a PhD thesis work. After mapping, we obtain bam file with all the information (i.e. sequence position, mapping quality, depth of coverage etc.). In my opinion, this is the most important data file one can have from the sequence data. As from here, we can do any populations genetic analysis (for example it can be used for variant calling, as well as it can be utilized to even get the raw sequence reads that were used to map). So having bam files are enough to combine different projects as everything else can be done from here bioinformatically.

After mapping the sequence reads, it is time to do the variant calling. Performing a proper variant calling is the most important step of all the genetic analysis. As the similarity between humans is high, it is futile to keep all the positions of the genome to do analysis (except for some analysis like Pairwise Sequentially Markovian Coalescent [PSMC] or Multiple Sequentially Markovian Coalescent [MSMC]). Therefore, in this step, the monomorphic sites are removed and only the variable positions are kept in the particular data set. The main sources of error come from a dilemma to define a position as an invariant or variant. This is not so big problem for common variants, which are shared between multiple individuals, thus would be found in several individuals. The problem arises when we have low covered position and only one individual in the data set is heterozygous for that position (which is called singleton). This heterozygous Single Nucleotide Polymorphisms (SNP) can be a true heterozygous position or to be detected because of sequencing error. Still now, no perfect solution has been developed to solve this problem. We can remove all the singletons from the data set saying that it is because of sequencing error but then we are biasing against a true SNP and this way we would lose a very important fraction of genetic variation (Casals and Bertranpetit, 2012). Therefore, it is a dilemma between false positive and false negative. The best way to solve this kind of situation is to use the best tool available for variant calling. In this thesis, we used Genome Analysis ToolKit (GATK) (McKenna *et al.*, 2010), the best one available when we started our project (Liu *et al.*, 2013). Apparently, at the end of my PhD we realised discrepancy between different lab results, which we hypothesised to be caused by differences in pipeline of variant calling (Methods and Results Section 3.3). We asked the population genetics community to revisit this problem and find the best method to do variant calling. Nonetheless, we used "HaplotypeCaller" from GATK to do the variant calling for our project. HaplotypeCaller algorithm is interesting and claimed to be better than other available methods to do variant calling. Haplotypecaller would remap the sequences for a particular region to test if the SNPs are coming because of mutation or they are coming because of an indel (as indel in a short read might be wrongly mapped as multiple SNPs for that region - Figure 1.7.1.1). Thus Haplotypecaller is a better choice than using traditional variant caller as it has better power to detect indels as well as it has lesser false positive for SNPs which are created falsely because an indel which is wrongly mapped as multiple SNPs. GATK also implement a variant recalibration step, which is a nifty tool. In a simple term, variant recalibrator is a machine learning approach to define if a position has a SNP or not. It has a priori knowledge of huge load of already known SNPs (for humans), which it would try to find in the particular data set that has been used for variant calling. From that it would extract different statistics for already known SNPs (for example Depth, Mapping quality etc.) and make a distribution out of it specific for that data set. The idea is that already known SNPs have lesser chances of being false positives, thus giving better information about the distribution of different statistics. Then it would look for a particular SNP (both known and unknown) in the data set. Looking for both known and unknown is important as not to bias against the novel variants (we have to stress the point that all the statistics by Variant Recalibrator were created on the data itself, thus it would have less bias for known and unknown SNPs). If the particular SNP was within this distribution, it would pass the filter test but if the SNP was clearly out of this distribution, it would reject the SNP calling. For example, for all the known SNPs the distribution of coverage is 4x-20x for the particular data set (from 1 percentile to 99 percentile). Therefore, for a SNP the coverage suddenly is exhibited to be 100x which is clearly outside the expected value of coverage for a given SNP. Thus, it would fail the filter test. It is a nifty tool but not workable for nonhuman organisms right now as we do not have a good database for known SNPs for other organisms. Finally, there are various ways to test if the variant calling was done properly (like transition-transversion ratio which is constant for humans) but the best way to test it is by doing population genetic analysis on the data set and comparing with already known scientific results. Of course, that does not mean to throw out every single new result that it produced (then the thesis or science in general would be boring). Rather try to check what was wrongly implemented that can possibly cause this discrepancy and if possible try to rectify that. There is no shortcut in this step, as something wrong in this step would haunt later in the downstream analysis.

Figure 1.7.1.1: Mapping dilemma. Where an insertion is wrongly interpreted as 3 different SNPs


Insertion
Reference sequence
C A G C T G G A A A T T * * C C T T C T T T G A A G T C
Read supporting reference
C A G C T G G A A A T T * * C C T T C T T T G A A G T C
Read supporting insertion
C A G C T G G A A A T T G C C C T T C T T T G A A G T C
Misaligned read
G C * * C C T T C T T T G A A G T C
Misaligned read
C A G C T G G A A A T T * * G C C
False SNP calls

1.7.2 Principal Component analysis (PCA) and Admixture

Although I am describing PCA and Admixture together here, the underlying theory is thought to be distinct. Interestingly both of them convey more or less similar results when using SNP data, though they were proven to be essentially same both practically (Patterson, Price, *et al.*, 2006) and theoretically (Lawson *et al.*, 2011).

PCA is used in distinct scientific topics (population genetic studies, physics, statistics, medical science etc.) and fairly common to be used for different purposes. In a nutshell, when the data possess a lot of correlated variables, PCA would convert that to uncorrelated variables. These uncorrelated variables are called Principal Component (PC). Every individual possesses ~ six billion nucleotides in their genome (as we are diploid). Therefore, when we compare 10 individuals, we are left with ~ 60 billion nucleotides in total. Of course, the number of correlated variables is much smaller here, as only .1% of our genome is different from one to another. Therefore, we can assume that we are left with 60 million SNPs, which correspond to 60 million correlated variables (the real number is generally lower than that as those SNPs are also shared between individuals). Understandably searching for a pattern in this big data is complicated. Therefore, if we use PCA on this data set, it would produce fewer independent variables out of it. The interesting ones are denoted as the top ones, which describe most of the data (in this case the most number of SNPs showing similar pattern), and it becomes less important for higher PCs. We know most of our variations (in this case mutations), which are found in the human genome, are because of ancestry. Unlike bacteria most of our genome is thought to be non-functional or "junk". Therefore, most of the variations or mutations do not affect us and they would accumulate through our ancestry. Thus, most of the genetic differences between individuals are caused by our differences in ancestry, which in turn makes PCA a good tool to understand ancestry. The main problem with PCA analysis is that sometimes the PCs do not have any biological meaning, which means we should be cautious when making inferences from PCA (i.e. the pattern can be caused by the difference in sequencing method, differences in coverage, origin of sequence etc.).

On the other hand, Admixture uses a completely different tactic to find the ancestry. The basic idea is that every SNP should hold Hardy-Weinberg equilibrium within a population. If the population have a substructure, the Hardy-Weinberg equilibrium would not be maintained for those SNPs. So admixture (or structure) algorithm would try to search for the optimum number of substructure in every SNP and in the end, it would take an average of all the SNPs. This method is extremely good at detecting admixed individuals and the ancestry from which these individuals were derived.

1.7.3: D-stats and Treemix

The publication of Neanderthal genome in 2010 (Green *et al.*, 2010) completely changed population genetics studies. Although the introgression from Neanderthals to OOA populations is ground breaking enough, they have also developed a new method, which was called ABBA-BABA test (later named as D-statistics or D-stats). D-stats completely changed how we do population genetic studies. D-stats is a powerful tool and super robust but also difficult to understand. In general, it is unaffected by samples used, coverage, filtering options, effective population size or sample size (although we find that it might have some biases which we discussed in details in Methods and Results Section 3.2-3.4). The basic idea is based on incomplete lineage sorting. The Figure 1.7.3.1 gives an idea about the null hypothesis. The outgroup population (here Y is an out-group of both W and X) should share more or less similar number of derived alleles in total with in-group populations because of incomplete lineage sorting. So if we count ADDA (where X and Y share derived allele) and DADA (where W and Y share derived allele) combinations our expectation would be more or less having a similar number in total for the whole genome. To normalize the count, we generally divide that with the total number of such events found in that particular calculation (DADA+ADDA).

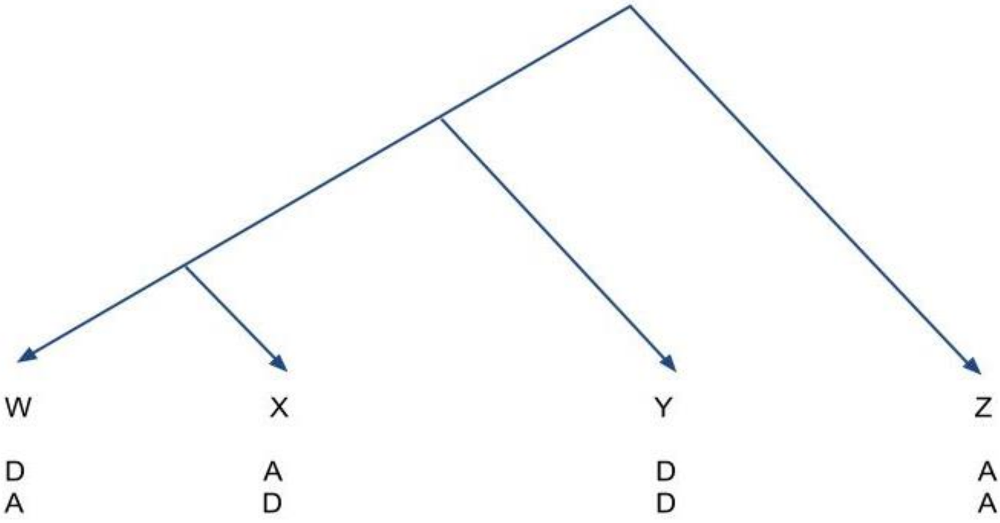
Null Hypothesis:

$DADA \approx ADDA$

$DADA - ADDA \approx 0$

$$\frac{DADA - ADDA}{DADA + ADDA} \approx 0$$

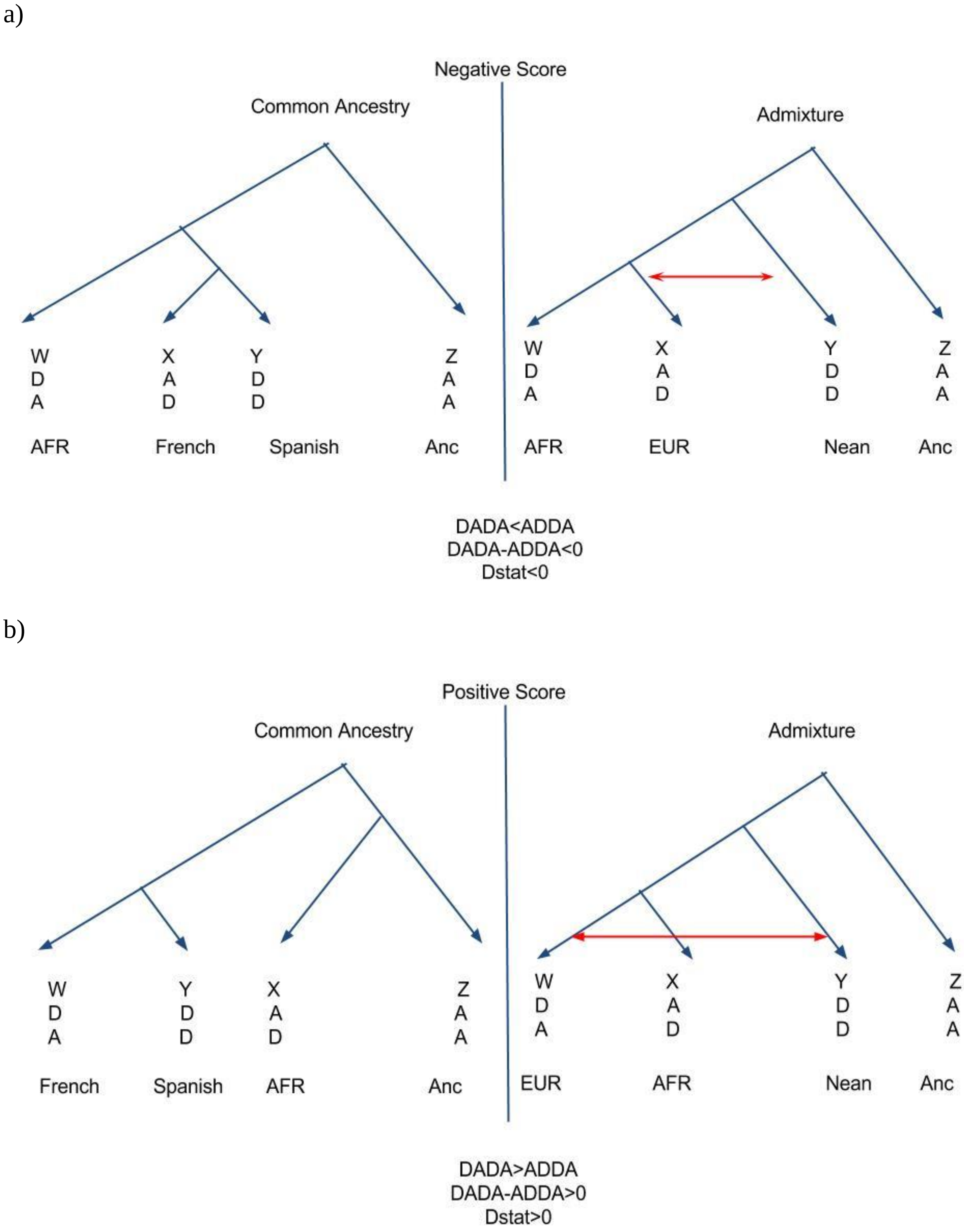
Figure 1.7.3.1: Null hypothesis for D-stats. W, X, Y and Z are 4 populations which are related with each other in the tree like structure which is drawn here. A stands for Ancestral allele and D stands for derived allele.



$$D_{stat} = \frac{(DADA-ADDA)}{(DADA+ADDA)}$$

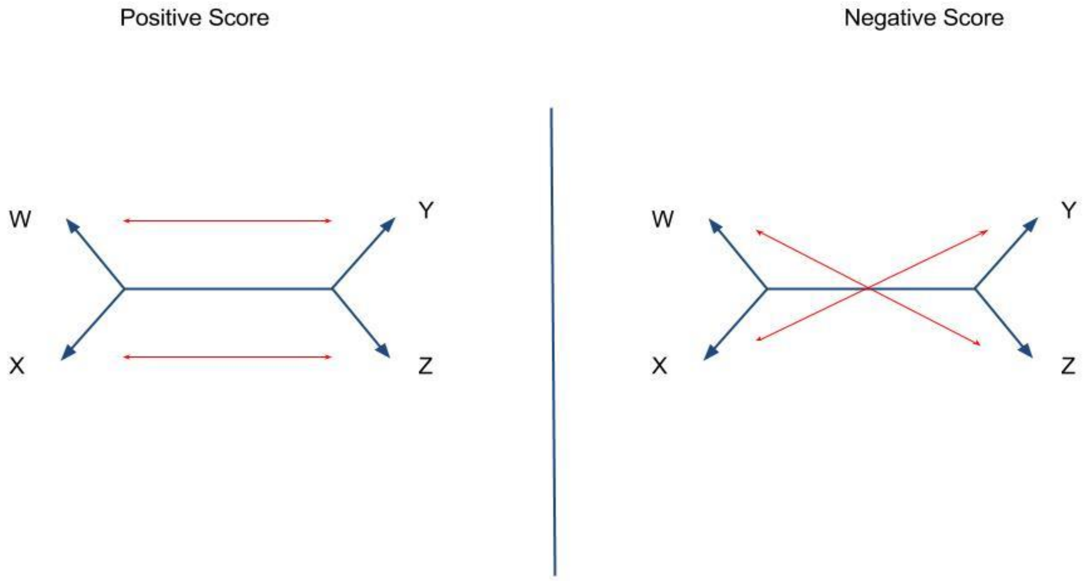
It suggests interesting results when it does not match with null hypothesis (when it gives a significant positive or negative result). Positive vs negative result is not interesting in itself as they only signify the direction. We can change the direction and thus interchange between positive and negative result of D-stats by interchanging between in-group (W and X). The real interest comes when it is deviated from zero (statistically significant). Here, the statistical significance is calculated by jack knife method. The idea behind calculating statistical significance from single individual is described in the next paragraph. Back to interpreting the D-stats results, the deviation from zero can be explained by two different demographic events. One is recent common ancestry and second is the admixture (Figure 1.7.3.2). Both of which can increase the derived allele sharing between one in-group and the outgroup population (W-Y or X-Y) compared to other combination (X-Y or W-Y). Although recent common ancestry can also be detected by other methods (Fst, simple genetic distances, MSMC etc.), the admixture event, that is detected by D-stats here, can only be detected by using this method. Admixture analysis, which we talked earlier, can also detect it but D-stats is much more powerful to detect those admixtures, especially admixture from a long time ago.

Figure 1.7.3.2: Alternative hypothesis for D-stats. W, X, Y and Z are 4 populations which are related in the tree like structure which is drawn here. A for Ancestral allele and D for derived allele. a) Showing explanation for positive score and b) showing explanation for negative score.



We can obtain statistically significant results from only one individual per populations, which might be contradictory to common sense. However, as our autosomes go through recombination in every generation, a single individual can produce a significant result by using clever tactics. The general idea is that by breaking the whole genome in multiple independent parts (i.e. Linkage Disequilibrium [LD] blocks) and calculating D-stats independently on them using Jack-knife method. In this method, we would remove a LD block at a time and calculate the D-stats for rest of the genome. Therefore, if we break the whole genome in 500 LD blocks, we would remove one block at a time and calculate the D-stats using the rest of 499 blocks. In this way, we will obtain 500 independent D-stats results from one individual. We can calculate an average as well as a standard deviation from these 500 independent D-stats. Jack-knife method can be used here, as in principle; every LD block would be independent of each other because we go through recombination in every generation. So here, one individual is not one data point, rather culmination of multiple independent LD block. The pattern of LD blocks is dependent on how our ancestors mated between themselves. Thus, we can have a statistically significant result from one individual. Although D-stats was used on a single individual by counting alleles, now the calculation is updated using allele frequencies (Patterson *et al.*, 2012; Meyer *et al.*, 2012). Thus, we can use it for multiple individuals now. Multiple individuals would give lesser standard deviation in principle. The direction of gene flow can be further simplified with the new update, as we do not need to use rooted tree any more (like Figure 1.7.3.2). Rather we can use an unrooted tree and from that understand the direction of gene flow (Figure 1.7.3.3).

Figure 1.7.3.3: D statistics representation in unrooted tree format



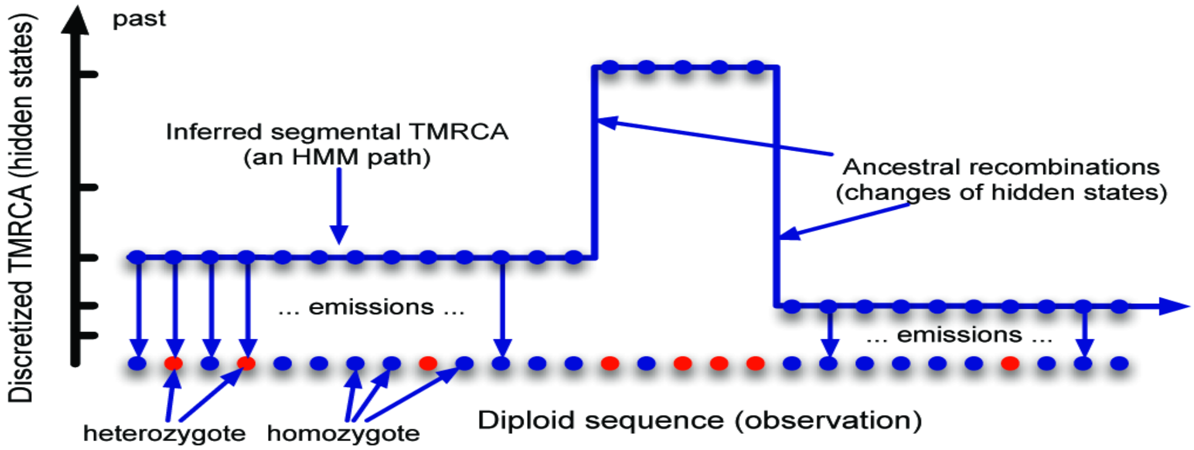
Treemix (Pickrell and Pritchard, 2012) although is not exactly same but essentially produces similar results. The basic idea is to obtain genome-wide allele frequency data according to populations. After that it would create a bifurcating tree using a Gaussian approximation to genetic drift (in a nutshell calculating Fst or drift like statistics between all the populations based on allele frequency and then try to build a tree using maximum likelihood method) using a known outgroup. However, when this approximation of genetic drift does not match well enough, it would invoke a migration from the most possible source population to most deviated population to fit the data better. Although it is easy to implement and gives interesting results, sometimes it can produce a wrong interpretation of the data. As this best method is based on maximum likelihood, sometime it can overfit the data depending on populations present in the data set. So the migration events shown by this method is not necessary to be true all the time (e.g. if Neanderthals sequences are not present in the data set it can show a migration from Chimp to OOA populations which is clearly wrong) but it can recommend a deeper understanding of the data by using different methods.

1.7.4 Pairwise Sequentially Markovian Coalescent (PSMC) and Multiple Sequentially Markovian Coalescent (MSMC)

Both PSMC (Li and Durbin, 2012) and MSMC (Schiffels and Durbin, 2014) are fascinating population genetics analysis tools. Both of these methods shared same basic concept. PSMC can be used for a single individual (a pair of chromosomes from a single individual) but MSMC can be used for multiple individuals. MSMC, of course, being the newer one uses better algorithm and concept to predict true demographic events than PSMC. MSMC is more efficient, can be used for multiple individuals and if phased can be used for calculating population separation time.

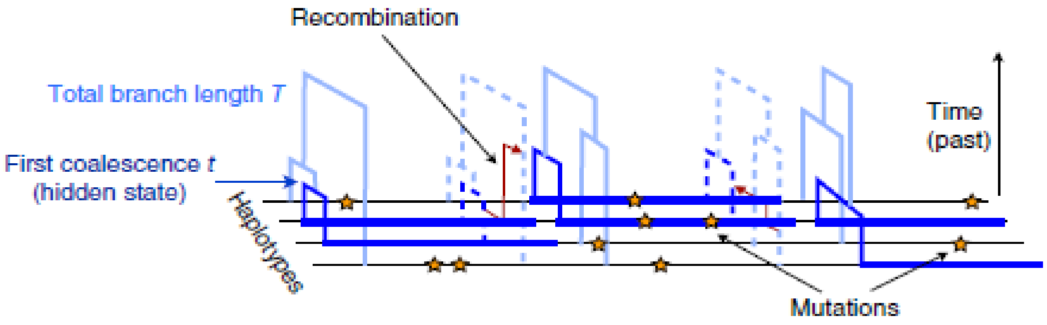
The Most Recent Common Ancestor (TMRCA) is useful to calculate history of effective population size changes. This is straightforward in the case of Mitochondrial DNA and Y chromosome as they do not recombine but (as already discussed) both Mitochondrial DNA and Y chromosome have TMRCA around 200 kya, this method (TMRCA) fails to give information beyond that time point for modern humans. On the other hand, autosomes hold information from much older TMRCA. Autosomes recombine in every generation, thus break up the genome in smaller part. These smaller parts act as independent loci, which do not necessarily have same TMRCA. If all the ancestral recombination events were known, calculating TMRCA for independent regions (flagged by two recombination events) would be straightforward. Sadly, we do not know where the recombination events happened thus, there is no way to detect independent region to calculate TMRCA. One way these two methods solved this dilemma is by looking for the average number of heterozygosity or segregating sites for a given length. The idea is if an ancestral recombination event happened, we should see a sudden change in the amount of heterozygosity or segregating sites for that region (Figure 1.7.4.1). The idea is if adjoining regions have similar number of heterozygosity (in case for diploid) or segregating sites (multiple individuals) the TMRCA for that region is also similar (as mutation occurrence is linearly related with TMRCA) but adjoining regions having differences in heterozygosity or segregating sites suggests differences in TMRCA which might be caused by ancestral recombination event causing two different TMRCA region come close to each other. So thus they would look for a sudden change in heterozygosity (PSMC) or segregating sites (MSMC) for a region thus define that region had an ancestral recombination event. After getting all the past recombination events, we can take independent regions (flagged by ancestral recombination events) and calculate TMRCA for that regions. After calculating TMRCA distribution of every such region, these programs use Expectation–Maximization (EM) algorithm to find maximum likelihood for the history of effective population size changes.

Figure 1.7.4.1: Schematics of how PSMC detects Ancestral recombination (Heng Li et al. 2012)



MSMC is similar to PSMC but can be used for multiple individuals. When using multiple individuals, we can have multiple TMRCA for a single region. MSMC would assume that for a region if there are multiple TMRCA, the lowest one is true for that region. MSMC has the upper hand as it can be used to calculate the divergence time between two individuals (thus two populations if those two individuals came from different populations). This cannot be done directly in case of PSMC as it can only use one individual at a time (although by some trick we could feed PSMC the haplotypes from different individuals thus creating a chimeric individual to calculate divergence time). Divergence is calculated simply by estimating effective population size for a given time for all the individuals together and then calculating effective population size for those individuals separately. The idea is that if they have not separated yet, effective population size within and across populations would be more or less same. Thus, the ratio would be close to one. However, if they have separated from each other for that time point, the effective population size within the population would be much lower than across populations (in principle after complete split effective population size across population should reach infinity) thus the ratio would be close to zero.

Figure 1.7.4.2: Schematics of how MSMC calculated TMRCA (Schiffels et al. 2013)



1.7.5 Simulations and dadi

Simulation models are indispensable for population genetic studies nowadays. Simulations can be used to sometimes get hidden parameters or demographic events as well as can be used to assess significance in selection tests. Both of them were used in our result sections extensively (Methods and Results Section 3.2-3.5). Hardy-Weinberg (Hardy, 1908) proved that if a population is infinite, pan mixing, with no migration, mutation or selection happening and have non-overlapping generation time; the allele frequency of this population would not change with time. One of the big assumptions here is having an infinite population size, which is generally not true. So Wright-Fisher model (Fisher, 1930; Wright, 1931) put the concept of effective population size thus making infinite population size finite. Now being finite population allele frequencies can have drift, which we can calculate from Wright-Fisher model. We can effectively simulate any population with some improvements in the Wright-Fisher model (like migration, mutation, selection etc.).

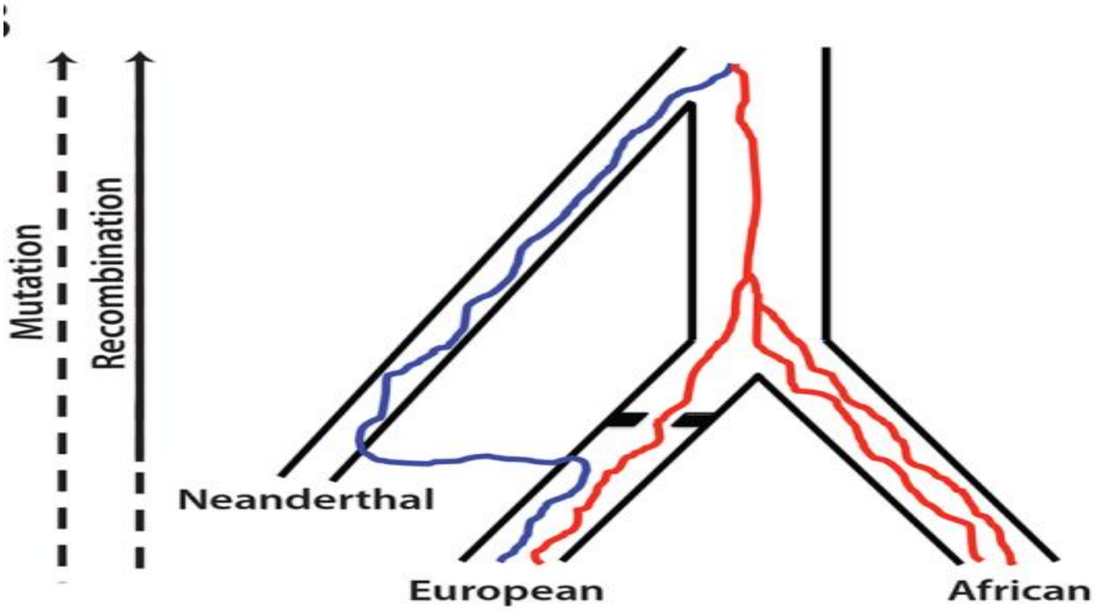
These simulation models have demographic parameters, which are not easy to calculate. If the population is well known (i.e. European, African or East Asians), we can use already known such simulation models (Gravel *et al.*, 2011). However, if the population is less characterized, we have to build the model ourselves. This can be done by dadi (Gutenkunst *et al.*, 2009). The basic idea is that we can build a model from allele frequency spectrum of different populations.

It uses multiple machine learning approach to find population demographic parameters. However, like any machine learning approach, it has a problem of overfitting. If prior knowledge of the demographic event is unavailable, it can fit impossible demographic parameters to match with empirical data. Thus, these kinds of methods are not useful for finding demographic events, which shaped the populations, rather should be used to fine-tune those demographic parameters, which have obtained by other methods (i.e. Treemix, MSMC etc.). We used dadi to fine tune of our already known demographic events of Andamanese (the basic demographic events were already obtained from other methods) to get those parameters (Methods and Results Section 3.2).

1.7.6 S* and D-stats by position

S* (Vernot *et al.*, 2014) is used for detecting introgressed regions in modern humans. The basic idea is that if a region is introgressed from hominin to modern humans, these regions would have high divergence time as well as would have long haplotype block as it recently introgressed in modern humans and recombination have not enough time to break it up properly (Figure 1.7.6.1). It needs a null model (model without introgression) to find introgressed regions. Any region, which is more than 99 percentile of the null distribution, is defined as an introgressed region. This null model also makes it hard to implement, as not every population has a null model. In addition, it has high rate of false positives.

Figure 1.7.6.1: Schematic of hominin introgressed regions (Vernot et al. 2014)



On the other hand, D-stats by position (Methods and Results Section 3.2, 3.3 and 3.5) is relatively easy to implement. The basic idea is that if a region is introgressed from hominin population to OOA populations, it should not be present in African populations. Therefore, we looked for regions, which lacked African-derived alleles in OOA populations using D-stats. This method is easier to implement, as we do not need to implement any simulations. It has a high false positive rate like S* (might be lowered by using more individuals). As both S* and D-stats by position both have a high false positive rate, we decided to put both of them together (if a region is positive for both S* and D-stats by position, we would define it as an introgressed region). It looks like if we do that the false positive decreases a lot and we can correctly extract those hominin introgressed regions (Methods and Results Section 3.2, 3.3 and 3.5).

1.7.7 Randomized Axelerated Maximum Likelihood (RAxML)

RAxML (Stamatakis, 2014) is a tool to create phylogenies from large data sets. RAXML specifically can handle large data sets. Of course, it cannot handle unphased autosomal data (recombination problem), but it is good for Y chromosome and Mitochondrial data as they do not recombine. The main problem with any phylogenetic study is that it can be computationally expensive, as possible topological spaces increase exponentially with the number of haplotypes (there are more than 10 million combinations for an unrooted tree with 10 haplotypes). We need to use some tactics to tackle this problem. All the algorithms for creating phylogenetic trees use clever tactics so that they do not look for every possible combination (RAXML uses parsimony tree approach to tackle this problem), rather they would use most likely combinations thus reducing topological spaces and can be done relatively faster (trade-off between robustness and time). This algorithm might not give the best tree possible but generally, they predict close to the real tree. Here in this thesis we used this programme to create the phylogenetic tree for Y-chromosomes and gave a new insight of Y-chromosomal distribution of Indian populations (Methods and Results Section 3.6).

1.8 Natural selection shaping modern human populations

After knowing the demographic events, it is now time to understand how natural selection has shaped human populations. Although modern humans have a recent origin (not more than 200 kya), we can observe ample amount of phenotypic differences between different populations. One way to explain these phenotypic differences is because of natural selection occurring in these populations. Especially if modern humans have left Africa recently (around 70 kya), all the OOA populations faced a completely different environment than they were adapted by living in Africa for thousands of years. Indeed, in our lab previously it was found that OOA populations have a higher amount of positive selection signature on their genome compared to

African populations (Pybus *et al.*, 2015).

In the late 60s, a big debate was brewing about the prevalence of natural selection, as Kimura introduced the neutral theory of molecular evolution (Kimura, 1968) which was a stark contrast to the accepted view of selection on that time (selectionist view). Fast forward to 2016, we have accepted that most of the human genome is not under selection, rather they behave neutrally (which also helps us to understand demographic events), but that does not necessarily mean that natural selection does not happen inside our genome. It is just that the natural selection happening to few portions of our genome, which is important for us (other parts are free to evolve neutrally). Natural selection can be divided into four broad categories: Purifying, Positive, Balancing and Sexual selection.

Purifying Selection: In my opinion, purifying selection is the strongest one. In this process, new deleterious mutations are removed from a population in every generation. We have four billion years' worth of knowledge by the natural selection process, which makes us efficient to survive in our environment. Of course, any diversion from this four billion years' worth of efficient knowledge most likely would put us in non-optimum space. For example, eukaryotic DNA replication machinery does one mistake for every one billion nucleotides addition (McCulloch and Kunkel, 2008). Any slight diversion (through random mutation) from this efficient method would put that organism to apparent disadvantage to his peers and thus would be eliminated from the gene pool.

Positive Selection: Although we have four billion years of knowledge, that does not mean that no improvement is possible. Positive selection is the opening for the improvements on what is already known. It is also a way to adapt in a constantly changing environment. For example, having digestive enzyme turned on to digest milk in adulthood does not make sense. Thus, most mammals turn off this enzyme when they become adult. However, as humans learn to domesticate cows having a mutation, which can turn on milk digestion in the adulthood, become beneficial and have shown to be positively selected in case for modern humans 8 kya. As Darwin put "It is not the strongest of the species that survives, nor the most intelligent that survives. It is the one that is most adaptable to change".

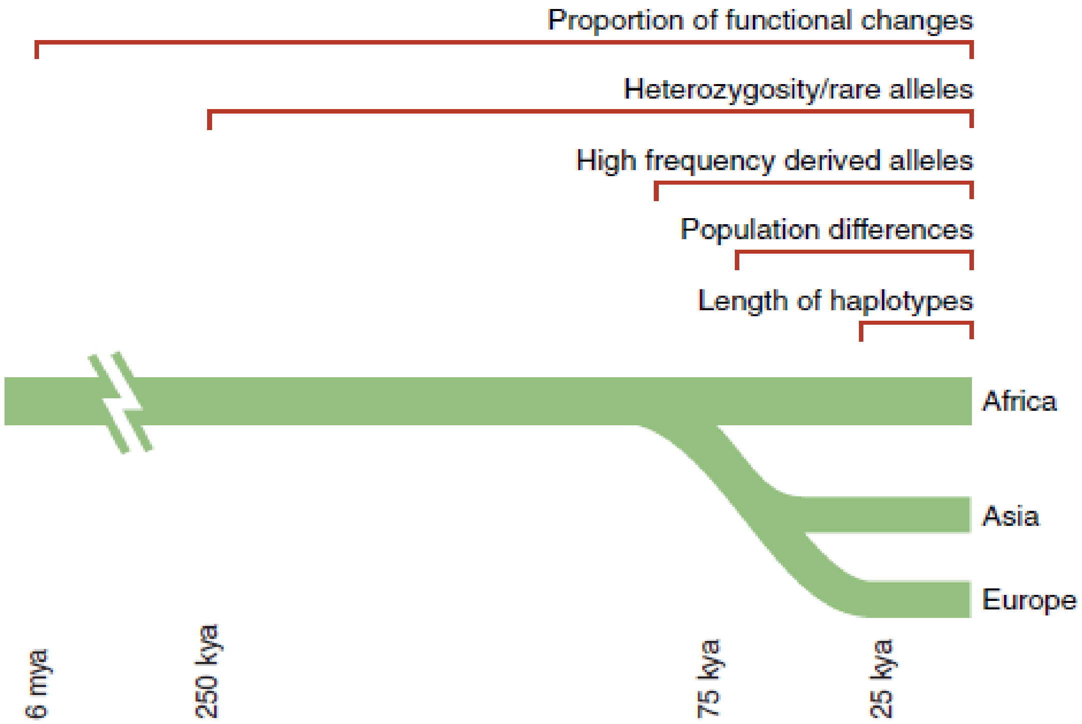
Balancing Selection: Sometimes variation is important to survive. When no clear better option is present, selection would maintain two different alleles in a population. This is more prevalent in the case of immune-related genes. We are in constant arms race with our parasites [a la Red Queen theory (Van Valen, 1973)]. As it is always difficult for parasites to develop resistance when more than one immune-related variations are present in a population. Sometime heterozygotes are also more advantageous than both of the homozygote. In that case, we can also see balancing selection happening to the population. One of the best examples is found in

African populations. In Africa malaria is prevailed which can cause a large number of individual to die early. It was found that some individuals have mutation to produce sickle cell Red Blood Cell (RBC). It was found that having a normal RBC would cause malaria pathogen to grow and kill the individual, whereas having homozygote position for this sickle cell mutation was not viable also as individual having this homozygotic mutation would not survive due to deformed RBC. However, having both mutations (i.e. for a normal RBC as well as for sickle cell mutation) in an individual would increase the chance of survival of that individual compared to both homozygotic individuals. Thus balancing selection would try to keep both of the mutations in the population.

Sexual Selection: Sexual selection occurs when the mating choice prevents random mating scenario and prefers one to another. Sometimes this is caused by specific phenotype, which is indeed better than the other one. Sometimes it can be a random choice of mating partner preference [although right now Handicap theory contradict that (Zahavi, 1975)]. One of the best examples of sexual selection is found in male peacock's tail size. The male peacocks have a big and grandiose tail, which put them at greater danger from predators as they can catch them more easily than the female peacocks. Although this can be disadvantageous for survival, this characteristic is highly regarded or valued by female peacock. Thus, this apparent disadvantageous trait has been selected through sexual selection.

In this thesis, we concentrated mainly on positive selection (especially hard sweep model). Various ways have been developed to detect positive selection (Figure 1.8.1). As all modern humans have relatively new origin, divergence data is not useful to detect positive selection within different populations of modern humans. Therefore, we concentrated on polymorphism data. Selection tests based on polymorphisms data can be divided into three broad classes.

Figure 1.8.1: Time scale for detection of Natural selection (Sabeti *et al.*, 2006).


Proportion of functional changes
Heterozygosity/rare alleles
High frequency derived alleles
Population differences
Length of haplotypes
Africa
Asia
Europe
6 mya
250 kya
75 kya
25 kya

Tests Based on Site Frequency Spectrum: The idea is that if a selective sweep had happened, it would distort site frequency spectrum around that region. Because of selective sweeps hitchhiking effect, the haplotype carrying selected allele would rise in frequency. This phenomenon would cause lower diversity, an excess of rare and derived alleles and absence of intermediate allele frequencies for that region. These changes of allele frequencies can be detected by various methods. In this thesis, we have used Tajima's D (Tajima, 1989), CLR (Nielsen *et al.*, 2005), Fay and Wu's H (Fay and Wu, 2000), Fu and Li's D (Fu and Li, 1993).

Tests Based on Long Haplotypes: Selective sweep not only changes site frequency spectrum around selected regions, it also increases LD around that region. As selection would always favour the haplotype containing the selected allele, it would have less time to recombine than neutral regions. We have used XP-EHH (Sabeti *et al.*, 2007), ΔiHH (Voight *et al.*, 2006), iHS (Voight *et al.*, 2006) and EHH average (Sabeti *et al.*, 2002).

Tests Based on Population Differentiations: If due to differences in the environment, an allele is selected in one population but the selection pressure is absent in other populations, it can be detected by the differences in allele frequency between those populations. We have used Fst (Weir and Hill, 2002) for our genotype data study (Methods and Results Section 3.1), although Fst could not be used on our whole genome sequence data (Methods and Results Section 3.2) because of low sample size. XP-EHH in a sense is also a test based on population differentiation. This is a merge between long haplotype-based test and population differentiation test together.

As there exist various methods to detect similar events (positive selection sweep), a consensus method was needed. We used Hierarchical Boosting (HB) strategy to detect hard positive selective sweeps (Pybus *et al.*, 2015). In this method, we have simulated data of null model (without selection) and hard selective sweep model. By comparing these two data sets, we could build a model, which can predict the best way to detect selective sweeps given multiple selection test results. Therefore, in a nutshell, it's a way to give a composite selection test score which is calibrated using a simulated model. We found it is robust to demographic changes of populations (at least for OOA populations).

2. Objectives

The main objective of this thesis is to describe **how demography and selection have shaped Indian populations**. Given the size and geographic position of Indian populations, it is easy to contemplate the complication of Indian populations' history. Indian populations with more than one billion people are underrepresented in genomic studies until now. Although 1000 Genome and other studies have sequenced many Indian individuals, all of them have focused on non-tribal populations of India. Tribal populations of India can give us the idea of how non-tribal populations are originated. As India is thought to be one of the oldest occupying places on earth for modern humans, they are also very important to know how modern humans spread Out of Africa. Before starting of our thesis, it was known that India has a complex ancestry and one of the main clines of ancestry is north to south. Northern populations has higher amount Ancestral North Indian (ANI) component and South has higher component of Ancestral South Indian (ASI) component (Reich *et al.*, 2009). Andamanese is undoubtedly the most interesting population from India given that their distinctive so-called “Negrito” morphology (Huxley, 1870) and the unclassifiable language they speak (Abbi, 2009). Their ancestry is also controversial. Morphological studies regarded them as a remnant of first out of Africa dispersal event and genotype data points out that they are similar to other Asian populations and originated together. However, this genetic hypothesis also raises the question why Andamanese looks so different than other contemporary Asian populations? We tried to address all these issues in this thesis.

The specific objectives of this study are:

- Elucidating if this cline exists. In addition, which populations would be best representative of these ANI and ASI components? This question will also indirectly address what is the relation of Indian populations compared to other reference populations (Africans, Europeans and East Asians) and thus will solve portability problem of GWAS studies from other populations to Indian populations.
- Analysing the Andamanese ancestry from a population genetic point of view. Trying to answer if they are remnant of first out of Africa event or they are related with other contemporary Asian populations.
- Characterizing regions under strong natural selection of Indian populations.

The first problem was answered in three different chapters:

1. A preliminary analysis using Genotype and Exome data of 120 Indian individuals from two different non-tribal populations (Methods and Results Section 3.1).
2. Whole genome sequences of 70 Indian individuals (Methods and Results Section 3.2).
3. Y chromosome analysis of Indian populations (Methods and Results Section 3.6).

The second objective was addressed in one chapter (Methods and Results Section 3.2) and the

third objective was addressed in two chapters (Methods and Results Section 3.1 and 3.2).
We unexpectedly found a hominin introgression in Andamanese instead of finding first out of Africa admixture. We changed our objective slightly at the later part of my thesis to understand the ancestry and the extent of introgression of this unknown hominin population. In our second paper, we only concentrated on Andamanese. As this is a big discovery we re-examined our introgression hypothesis (Methods and Results Section 3.3) and then restudied some other populations to know the extent of affected populations who have introgression from this unknown extinct hominin population (Methods and Results Section 3.4 and 3.5).

3. Methods and Results

3.1 Population and genomic lessons from genetic analysis of two Indian populations.

Juyal G, Mondal M, Luisi P, Laayouni H, Sood A, Midha V, et al. Population and genomic lessons from genetic analysis of two Indian populations. Hum Genet. 2014 Oct 1;133(10):1273–87. DOI: 10.1007/s00439-014-1462-0

3.2 Genomic analysis of Andamanese provides insights into ancienthuman migration into Asia and adaptation.

Mondal M, Casals F, Xu T, Dall'Olio GM, Pybus M, Netea MG, et al. Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. Nat Genet. 2016 Sep 25;48(9):1066–70. DOI: 10.1038/ng.3621

3.3 Further confirmation for unknown archaic ancestry in Andamanand South Asia.

Mondal M, Casals F, Majumder PP, Bertranpetit J. Further confirmation for unknown archaic ancestry in Andaman and South Asia. bioRxiv. 2016 Aug 23;71175. DOI: 10.1101/071175

3.4 East Asian introgression: an overall vision of ancient introgressions in the human lineage.

Mayukh Mondal et al.

Manuscript Under Preparation

Abstract

We have recently published a paper (Mondal et al. 2016) showing that an unknown archaic population (an extinct Eurasian hominin which is neither Neanderthal nor Denisova) introgressed in the Andamanese and Indian populations. In this work we try to delimit which are the populations having had introgression from this unknown hominin. By using whole-genome sequences available and comparing to simulated data we hypothesized that Out of Africa (OOA) Eurasian populations of modern human have had introgression at least three times from extinct hominins. First from Neanderthal to all OOA populations, second all Asian and Pacific populations have had introgression from this unknown population and third, all Pacific populations having introgression from Denisova. In this follow up study we try to delimitate the geographic extent of the introgression with the unknown hominin population, which could be wider than initially proposed and encompass all Asian and Pacific populations.

Introduction

All modern humans are related with each other at a time depth of up to 200,000 years (Soares, 2009, Poznik et al. 2006, Li et al. 2011). Recently several papers from genetic studies have argued that there was only one OOA event (Mondal et al. 2016, Malaspinas et al. 2016, Mallick et al. 2016) that happened less than 100 kilo years ago (kya), contradicting an earlier hypothesis of multiple OOA event for modern humans (Rasmussen et al 2011, Grun et al. 2005). There are much older human remains in Eurasia previous to the modern human expansion, which are referred here as extinct Eurasian hominins; the ancestors of these hominins had left Africa much earlier than modern humans and there is no consensus on where the common ancestor of all hominins originated. Sequencing of ancient bones from Neanderthal and Denisova individuals led to the accepted introgression scenario (Green et al. 2010, Meyer et al. 2012), refuting an earlier hypothesis of no admixture between extinct Eurasian hominins and modern humans.

Recently we have argued that there was another extinct Eurasian hominin population, out-group of both Neanderthal and Denisova, which is found to be introgressed in the Andamanese populations (Mondal et al. 2016). We showed that this population might have introgressed in an area that would comprise Continental India, Tibeto-Burman speakers of East India (a recent independent work has also described the presence of introgression in Tibetan populations (Lu et al. 2016)), as well as Pacific populations, although the exact limits of this were not assessed. In the present analysis we attempt to clarify which modern populations show introgression by the same ancient hominin other than the Andamanese, by using the D-stats and F4 ratio tests D-stats are used to detect derived (D) alleles shared between two populations, compared to other population given an out group of all these three population. For example, when using D (African, European, Neanderthal, Ancestral) we analyse the amount of derived alleles shared between European and Neanderthal compared to African and Neanderthal. A significant amount of divergence from 0 suggests that an introgression happened in the European or African branch from Neanderthal. The sign of the value determines in which branch this introgression have happened (e.g. negative value in D (African, European, Neanderthal, Ancestral) suggest an introgression happened in European branch rather than African branch). The F4 ratio test is similar to D-stats, with the main difference being that F4 ratio test uses the ratio of two F4 or D-stats. While D-stats is known to be not appropriated to measure the amount of ancestry (Patterson et al. 2012), the F4 ratio test can perform that by using two different ratios of F4 (or D-stats). Importantly, this method is only correct if the assumed tree is true (Patterson et al. 2012).

Results

In our study on the Andamanese, East Asians displayed a slightly dearth of African derived alleles compared to Europeans (Figure 2 in Mondal et al. 2016). However, interestingly East Asian and Andamanese populations revealed to have higher amount of both Neanderthal and Denisova introgression (also or Indians) using D-stats test (Table 1a). This increased amount of Neanderthal and Denisova in Asians has been previously reported (Prüfer et al. 2014 Table S14.6, Meyer et al. 2012 Table S24 and S25 and Mondal et al. 2016). Interestingly when using the F4 ratio statistics using two Neanderthals, this increase was nullified and it was within just one standard deviation (Prüfer et al. 2014 Table S14.7 and S14.8). This discrepancy between D-stats and F4 ratio remained unexplained. The hypothesis is that East Asian got introgression from both Neanderthal and Denisova independently is unlikely (as well as for continental Indian and Andamanese, showing similar results). The fact that the amount of Neanderthal and Denisova introgression is similar between East Asians, Andamanese and Tribal Indians, suggests this introgression happened before these populations have separated from each other. The time span for these two events (introgression from both Neanderthal and Denisova) was small, making unlikely that these two extinct hominin populations lived close to each other and had made similar introgression patterns for all these diverse modern human populations.

One alternative explanation is that East Asians had some introgression from this unknown extinct hominin population also. Having introgression from any extinct Eurasian hominin population which is an out-group of Neanderthal and Denisova would increase the D-stats values of Neanderthal and Denisova, as this unknown archaic population shared an ancestry with them. In contrast, it would not be detected in the F4 ratio test using two different Neanderthals since this test, as our simulation models shows (Table 2b) it would be unaffected by the introgression from Unknown hominin population. We simulated this scenario and reached the conclusion that, as we hypothesized, if this introgression occurred, East-Asian and Andamanese would show more Neanderthal and Denisova ancestry compared to Europeans when using D-stats (Table 2a) but the increase would be insignificant when using F4 ratio test for East Asians (Table 2b) (that is a good approximation of what was found in real data by Prüfer et al. 2014 and Meyer et al. 2012). Interestingly under this model, Asian populations should show a 2% dearth of African alleles (Table 1a) like Andamanese; we detected this lack of African ancestry in case of Indian and Andamanese populations (as well as one Tibeto-Burman population) but failed to detect it in case of East Asian populations (Mondal et al. 2016); new high coverage sequences with a homogeneous calling are needed to solve it.

The case of Papuans is also very interesting. Although Papuans are known to have introgression from Denisova, we found that Papuan has much more Neanderthal introgression than other OOA populations when using D-stats (Table 1a). It is also interesting to note that when calculating Denisova introgression in Papuan using F4 statistics, different amounts of Denisova ancestry were estimated when using different populations (African, European, East Asian) as reference in the F4 ratio test (Table 1b). But using simple simulation of the three introgression model (Figure 1) we were able to explain most of the empirical results (Table 2b).

Discussion

East Asian populations might have introgression from an unknown hominin population, as described in the case of the Andamanese (Mondal et al. 2016). According to that, we have updated our model of introgression of extinct hominins into the modern human lineages (Figure 1). If true, this model would explain some contradictory results. Like why Tibeto-Burman population, with similar ancestry to East Asians lacks a 2% of African alleles. Second, it also solves how seemingly unrelated populations in South and South-East Asia revealed similar amount of introgression. If all Asian populations have introgression from the extinct hominin, then it is much easier to explain the geographic distribution of the regions with the introgression than in our initial work (Mondal et al 2016). Finally, it also explains previous results where Asian populations showed higher amount of Neanderthal and Denisova introgression compared to Europeans (Prufer et al. 2014 and Meyer et al. 2012) which has generated a big debate till now. We would also like to emphasize that if there is an introgression in Asian populations from an out-group, it would increase both Neanderthal and Denisova more or less in a similar amount due to the similarity of both to the archaic hominin that would have generated by incomplete lineage sorting when using D-stats (which is the case for both our real data in Table 1a and simulation data in Table 2a). If the introgression happened either from Neanderthal or Denisova, the increase in D-stats would not be uniform. We think that the alternative scenario of two independent introgressions of similar amount in Asian populations from Neanderthal and Denisova happening after they had separated from the Europeans and before their separation is highly unlikely. Finally, this also explains the high divergence of Denisovan introgression to Papuan detected by F4 ratio test.

We are unable to explain why East Asian populations have not showed similar less amount of African allele in Mondal et al. 2016 like other Indian or Andamanese populations. One possible explanation would be a more complicated ancestry than we have anticipated (admixing with a population that have not introgressed with this unknown hominin population or having less amount of introgression) or some sample size bias. Having less individuals do affect D-stats, which specifically detects less African ancestry as it can be seen in Papuan and Aboriginal Australian results (Mondal et al 2016 Figure 2). Also, when using two Andamanese, it decreases the amount of dearth of African alleles (Mondal et al 2016 Under Review). The power of D-stats using a single individual could not be enough to detect a lack of ancestry (at least for African) This hypothesis needs to be tested.

Methods

Data: We used the previously published data from Meyer et al. 2012 and Mondal et al. 2016. We downloaded the Neanderthals (Green et al. 2010), Denisova (Meyer et al. 2012) from their respective sites. We converted the vcf files to plink format using vcftools (Danecek et al. 2011) and then added the Ancestral information from 1000 genome data site (1000 Genomes Project Consortium 2012). We removed all the SNPs missing information for any individual using --geno flag in plink-1.07 (Purcell et al. 2007) to remove lowly covered regions, which can bias the frequency estimation for populations. We also removed all the transitions as ancient genomes are prone to have more transitions due to DNA degradation. We used qpDstat from Admixtools-1.1 (Patterson et al. 2012) to calculate the D-stats and F4.

Simulations: For East-Asian introgression we used the following ms (Hudson 2002) command:

Mscode:
```
ms 16 300000 -I 7 2 2 2 2 2 4 2  -t 7.44 -r 7.74 10000 -n 1 2.20 -n 2 4.47 -n 3 6.53 -
g 2 101.69 -g 3 146.31 -m 1 2 1.49 -m 2 1 1.49 -m 1 3 .46 -m 3 1 .46 -m 2 3 1.85 -m 2 3 1.85
-es .02 4 .97 -ej .02 8 7 -ej 0.022 4 3 -es .025 3 .97 -ej .025 9 5 -ej .029 3 2 -en .029 2 .29 -em
.029 1 2 8.93 -em .029 2 1 8.93 -es .034 2 .975 -ej .034 10 6  -ej .087 2 1 -en .23 1 1 -ej .235
7 6 -ej .28 6 5 -ej .364 5 1
```

Where the populations are: Africans, Europeans, East Asians, Papuan, Unknown, Neanderthal and Denisova. We used previously published model of European, East Asian and Africans (Henn et al. 2015). We updated the model to use 1.5x10^{-8} mutation per bp per generation and 30 years for generation time was used as recently suggested (Scally et al. 2012). As D-stats is robust and not affected by the effective population size and time of admixture, we added the other populations with the model using simple parameters. Separation time between Papuan and East Asians are set to 40 kya. Separation time between modern humans and archaic hominins set as 650 kya. Separation time between Unknown and Neanderthal-Denisova set at 500 kya. Separation time between Neanderthal and Denisova is set to around 420 kya. Introgression of Neanderthal to modern humans is set to 60 kya and 2.5% introgression. Unknown to all Asia Pacific populations are set 45 kya and 3% introgression. Denisova to Papuan 35 kya and also having 3% introgression. Real timing of this demographic event might be different but D-stats are robust to such changes and would not affect the results (not shown). Andamanese were not simulated separately. As D-stats results are robust, simulated results for East Asians are exactly same with Andamanese and Indian Tribal populations (results are not shown).

References

Abecasis,G.R. *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **135**, 0–9.

Danecek,P. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Green,R.E. *et al.* (2010) A draft sequence of the Neandertal genome. *Science*, **328**, 710–722.

Grün,R. *et al.* (2005) U-series and ESR analyses of bones and teeth relating to the human burials from Skhul. *J. Hum. Evol.*, **49**, 316–334.

Henn,B.M. *et al.* (2015) Estimating the mutation load in human genomes. *Nat. Publ. Gr.*, **16**, 1–11.

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

Li,H. and Durbin,R. (2011) Inference of human population history from individual whole-genome sequences. *Nature*, **475**, 493–496.

Lu,D. *et al.* (2016) Ancestral Origins and Genetic History of Tibetan Highlanders. 1–15.

Malaspinas,A.-S. *et al.* (2016) The genomic history of Indigenous Australia. *Nature*, **160100007**, 1–152.

Mallick,S. *et al.* (2016) The Simons Genome Diversity Project : 300 genomes from 142 diverse populations. *Nature*.

Meyer,M. *et al.* (2012) A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*, **338**, 222–226.

Mondal,M., Casals,F., Majumder,P.P., *et al.* (2016) Further confirmation for unknown archaic ancestry in Andaman and South Asia . *bioRxiv*, 1–10.

Mondal,M., Casals,F., Xu,T., *et al.* (2016) Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. *Nat. Genet.*, 1–102.

Patterson,N. *et al.* (2012) Ancient admixture in human history. *Genetics*, **192**, 1065–1093.

Poznik,G.D. *et al.* (2016) Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat. Genet.*, **12**, 809–809.

Prüfer,K. *et al.* (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, **505**, 43–49.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

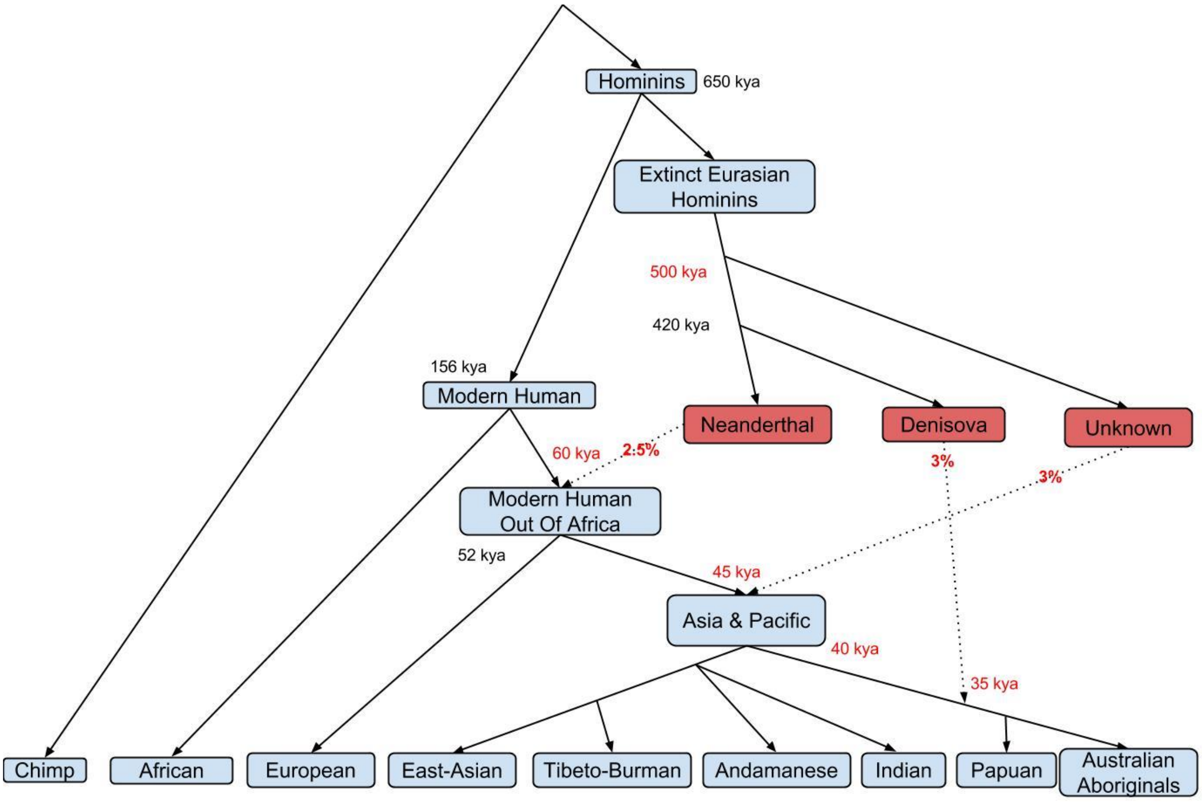
Rasmussen,M. *et al.* (2011) An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. *Science*, **334**, 94–98.

Scally,A. and Durbin,R. (2012) Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.*, **13**, 745–53.

Soares,P. *et al.* (2009) Correcting for Purifying Selection: An Improved Human Mitochondrial Molecular Clock. *Am. J. Hum. Genet.*, **84**, 740–759.

Figures

Figure 1: Model of Gene flow from archaic hominins to Out of Africa populations. Times of demographic events are written in kya (kilo years ago). Already known demographic events are written in black and assumed demographic events used in the simulation are written in red colour.



Tables

Table 1a: D-stats of Neanderthal, Denisova Ancestry. Where AFR=African (Yoruba, Mandenka, Mbuti), ASN=East Asian (Dai and Han), AND=Andamanese (Jarawa and Onge) EUR=Europeans (French and Sardinian).

W	X	Y	Z	D score	Z score
AFR	ASN	Neanderthal	Ancestral	-0.0458	-9.755
AFR	AND	Neanderthal	Ancestral	-0.0425	-10.322
AFR	EUR	Neanderthal	Ancestral	-0.0323	-8.507
EUR	ASN	Neanderthal	Ancestral	-0.0169	-3.196
EUR	AND	Neanderthal	Ancestral	-0.0127	-2.609
EUR	ASN	Denisova	Ancestral	-0.0121	-3.048
EUR	AND	Denisova	Ancestral	-0.0155	-4.168

Table 1b: F4 Ratio tests

f4	α	S.E.	Z
f4(NEN,Ancestra;PAP,AFR)/f4(NEN,Ancestral,DEN,AFR)	0.103687	0.009638	10.758
f4(NEN,Ancestra;PAP,AFR)/f4(NEN,Ancestral,DEN,EUR)	0.051148	0.009924	5.154
f4(NEN,Ancestra;PAP,AFR)/f4(NEN,Ancestral,DEN,ASN)	0.026386	0.010129	2.605

Table 2a: Simulation D-stats of Neanderthal and Denisova. Where AFR=Africa, ASN=East Asian, EUR=European, AND=Andamanese, NEN=Neanderthal and DEN=Denisova.

W	X	Y	Z	D score	95% confidence interval of Empirical data
AFR	EUR	NEN	Ancestral	-0.0318	-0.04 ~ -0.025
AFR	ASN	NEN	Ancestral	-0.0434	-0.055 ~ -0.036
AFR	PAP	NEN	Ancestral	-0.0558	-0.07 ~ -0.048
AFR	PAP	DEN	Ancestral	-0.0701	-0.068 ~ -0.048
EUR	ASN	NEN	Ancestral	-0.0139	-0.027 ~ -0.006
EUR	ASN	DEN	Ancestral	-0.0126	-0.02 ~ -0.004
EUR	PAP	NEN	Ancestral	-0.0287	-0.046 ~ -0.02
EUR	PAP	DEN	Ancestral	-0.0631	-0.083 ~ -0.059
ASN	PAP	NEN	Ancestral	-0.0162	-0.03 ~ -0.004
ASN	PAP	DEN	Ancestral	-0.0540	-0.075 ~ -0.049
EUR	ASN	AFR	Ancestral	0.0246	0.002 ~ 0.012
EUR	PAP	AFR	Ancestral	0.0496	0.022 ~ 0.035

Table 2b: Simulation result of F4 ratio test. Where AFR=Africa, ASN=East Asian, EUR=European, NEN=Neanderthal, NEN1= a second simulated Neanderthal and DEN=Denisova.

f4	**α**	**95% confidence interval of Empirical data**
f4(DEN,NEN;AFR,EUR)/f4(DEN,NEN;AFR,NEN1)	0.016	0.015 ~ 0.02
f4(DEN,NEN;AFR,ASN)/f4(DEN,NEN;AFR,NEN1)	0.018	0.016 ~ 0.02
f4(NEN,Ancestra;PAP,AFR)/f4(NEN,Ancestral,DEN,AFR)	0.105	0.084 ~ 0.124
f4(NEN,Ancestra;PAP,AFR)/f4(NEN,Ancestral,DEN,EUR)	0.050	0.031 ~ 0.071
f4(NEN,Ancestra;PAP,AFR)/f4(NEN,Ancestral,DEN,ASN)	0.03	0.06 0.047

3.5 Ancestry and unknown Hominin introgression in Australian genomes.

Mayukh Mondal et al.

Manuscript Under Preparation

Abstract

The demographic history of Aboriginal Australians has been largely uncharacterized until the very recent genome analysis (Malaspinas et al., 2016). This population was thought to be introgressed from Neanderthal (of around 2%, that could have originated about 60 kya) and Denisova (of around 4-5%, originating about 44 kya) (Prüfer et al., 2014). The introgression events may have taken place in several pulses, mainly the Neanderthal, affecting different populations at different depth. For example, Malaspinas et al. have hypothesised to have 5 different introgressions from two extinct Eurasian hominin populations (Neanderthal and Denisova) to three Out of Africa (OOA) modern human populations (Europeans, East Asians and Australians). The recognition of the introgression of archaic genomes is crucial to understand the origin of the populations. Thus, it is interesting to note that in Malaspinas et al., 2016 considering the introgression of extinct Hominin lineages allows distinguishing between single or multiple out of Africa scenarios. Similar results have also been found in other studies (Mondal et al., 2016; Mallick et al., 2016). Although recently another extinct hominin population has been hypothesized to have introgressed in Andamanese and other populations in South Asia (Mondal et al., 2016). Here we analyse if this introgression is also detected in Aboriginal Australian genomes, thanks to the available data and produce a complete picture of introgression by the various Hominin groups in the gene pool of Aboriginal Australians.

Our results show that indeed Aboriginal Australians have introgression from this unknown extinct hominin population which represents about 17 Mb (average length of 65kb and average amount detected 200kb) of the Aboriginal Australians genome in total.

Introduction

Archaeologists have found ~50 kilo years ago (kya) modern human remains in Australia. Although the time of first occupation of modern humans is disputed. it was earlier thought that Sahul was occupied around 40-45 kya (O'Connell and Allen, 2004). Recent studies supported much earlier date for Sahul occupation ~47.5-55 kya (Summerhayes et al., 2010; Clarkson et al., 2015; O'Connell and Allen, 2015). Coincidentally this time span overlaps with occupation of modern humans in Sunda region (Barker et al., 2007) and having similar morphological traits (Matsumura and Oxenham, 2014) suggesting a similar ancestor population occupying whole Sunda and Sahul regions around 50 kya.

A very early Australian settlement led to the hypothesis that Aboriginal Australians are remnants of the first Out Of Africa (OOA) expansion of modern humans. They also physically resemble Papuans and other SE Asian populations like Andamanese, Philippine and Malaysian negritos (all of them of very low stature and highly pigmented), populations that have had a strong isolation from surrounding ones and were considered to be related among them and the remains of a putative initial out of Africa wave of modern humans. Most of remaining Asians would have been part of a second out of Africa event. Even if this idea was initially confirmed with genetic data (Rasmussen et al., 2011) recent whole genome studies have given strong support to a common origin of all non African modern humans, and thus to a single out of Africa event (Mondal et al., 2016; Malaspinas et al., 2016; Mallick et al., 2016).

Interestingly Aboriginal Australians also harbour higher amounts of extinct Eurasian hominin

DNA due to recent introgression events (Meyer et al., 2012; Rasmussen et al., 2011; Malaspinas et al., 2016). Although the Denisovan introgression are proven to be right in Pacific populations, the exact amount of introgression is largely varied (3-6%) (Prüfer et al., 2014). Recently we have shown that Andamanese have introgression from an unknown hominin population with some other populations from South and Southeast Asia (Mondal et al., 2016). As this unknown hominin occupied in South Asia and Southeast Asia, they have higher chance of introgression in Australian populations. Here, we want to test if, in addition to Neanderthal and Denisovan, this population also had introgression in their genome from other unknown hominins, as described for the Andamanese populations (Mondal et al., 2016).

Results

A new Andamanese sequence

Besides the ten modern Andamanese individuals described in Mondal et al. (2016), a new sequence has been obtained from an ancient specimen. In all the analysis below, this new sequence has been used as Andamanese.

Total D-stats

We detected with D-stats a significant lack of African derived alleles in Australians, compared to Europeans or East Asians, thus suggesting that Australians also had introgression from the extinct hominin population. Australian and Papuan showed ~3% less African ancestry compared to Europeans (Figure 1). The new Andamanese sample also showed a lack of African ancestry although in lower proportion than Australians.

Treemix Analysis

We performed the Treemix analysis with whole genome sequences of Aboriginal Australians and Papuan with other already known key reference populations around the world (Initial SGDP data set Mallik et al (2016)), including two extinct groups, Neanderthal and Denisova (Green et al., 2010; Meyer et al., 2012). Results (Figure 2) show a clear clustering of modern humans, in which, after the African initial splits, the European separation leaves all Asian (and American) and Pacific populations in a clade, with Pacific populations (Aboriginal Australians and Papuan) making a cluster which was also reported earlier (Mondal et al., 2016; Mallick et al., 2016).

We then allowed for migrations in the Treemix analysis. The first migration is coming from Denisova to Pacific populations. The second migration is coming from Neanderthal to all Out of Africa (OOA) populations. The third migration is coming from Europeans to Australian aborigines. These three migrations have already been described (Meyer et al., 2012; Green et al., 2010; Malaspinas et al., 2016). It is interesting to note that although aboriginal Australian in the data set are known to be admixed with Europeans with important amount (9%), the first migration in Treemix was not coming from Europeans to Australians Aboriginals; this suggests that Treemix analysis is more powerful at detecting migration if the populations have diverged long time ago (in this case 600 kya for the hominins that introgressed), even if the admixture portion is much lower (~3-6%) (Prüfer et al., 2014). We did not detect any other hominin introgression in this analysis.

Detecting Introgressed Regions for Unknown Hominin

We then concentrated on the unknown archaic ancestry as detected for Andamanese (Mondal et al., 2016), which could be present in Aboriginal Australian genomes. In the case of Andamanese, as they have no Denisova ancestry, finding these unknown hominin introgressed regions are relatively straightforward. We have performed extensive analyses [D-stats (Patterson et al., 2012), D-stats by position (Mondal et al., 2016) and S* (Vernot et al., 2014)] to prove that Andamanese indeed have some introgression from this extinct population. We looked for regions which have absence of African allele compared to European or East-Asians. We extracted ~37 mb region which we then tested for introgressed region using S*. We have been able to extract close to 2.5 Mb of DNA of that putative origin that, as seen in Figure 3a, is a region close to the extinct hominins of comparison. This amount (below 1%) should be considered as the minimum amount of introgression, as this analysis has only considered the regions with high likelihood of being introgressed.

Then we concentrated on Aboriginal Australians. As they have introgression from Denisova we need to avoid the regions detected as introgressed. We masked any region having derived alleles which are shared between Denisova and Aboriginal Australians but absent in European and Africans using D-stats by position (see Methods). The idea is if a region introgressed from Denisova to Aboriginal Australians, it will have derived allele shared between these two populations but will be lacked in Europeans and Africans. So for that region D (Australian, European; Denisova, Ancestral) would give positive values and thus would not be used in S* in later step. Table 1 shows the amount of introgressed region, with first column using D-stats by position in the whole sequences, the second when masking for Denisova and the third, much more stringent, using the last and also S*. There are strong differences among individuals, mainly due to the amount of recent (and mainly European) introgression.

We then merged all the unknown hominin introgressed regions from the 83 Aboriginal Australian genomes together and obtained a total of 16.7 mega bases of high confidence introgressed region. Taking this region (in the cases of overlap between two or more individuals, just one is taken), we calculated distances among several modern genomes (African, European, and East Asian) and two ancient (Neanderthal and Denisova) and plotted a simple tree (Figure 3b). Result is clear: the genome of the proposed archaic hominin is close and basal to the extinct hominins.

It is interesting to note that the average length of such regions is ~65kb which is similar to the new Andamanese sample introgressed regions as well to the regions found in the ten modern Andamanese (65 Kb) (Mondal et al., 2016). All this suggests that introgression happened at the same time, most probably before the separation of the Australian branch.

Discussion

We found independently the unknown ancestral hominin population both in a new Andamanese genome and in Aboriginal Australians. It is interesting to note that if introgression happened in the ancestors of both, Andamanese and Australians, it had to happen before their divergence. Thus, it should be very basal at the diversification of Asians and Australians. This also suggested by same average length of these regions. The introgression had to happen after separation of the Europeans and before the diaspora towards

Australia. Due to the masking of Denisova genome, the amount of sequence that is unequivocally attributable to the unknown archaic ancestry in Australians is small and difficult to estimate the exact amount of introgression; only a lower bound may be established. The present results suggest that the introgression happened in a basal Asian (or South/ Southeast Asian) and thus the introgression would be more widespread than initially postulated, affecting also the East Asia populations which we also postulated earlier (Results 3.3).

Methods

Samples

We used whole genome sequences from eight different populations of 83 individuals of Aboriginal Australians (Malaspinas et al., 2016). We also used several individuals from different populations around the world as comparison purpose from Reich panel (Mallick et al., 2016) and some Papuan individuals. We have also used an ancient Andamanese individual for our analysis.

Data

We used the whole genome vcf file that was generated in the Australian study (Malaspinas et al., 2016).

Treemix

Treemix (Pickrell and Pritchard, 2012) was used to analyse the divergence of the populations from each other and admixture within populations, using the data described above. We used ancestral file from 1000 Genome (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/) (Abecasis et al., 2012). We used the -k flag to allow for LD. We defined LD blocks as 1 Mb in length, which in this study is corresponded to about 5,000 SNPs.

D-stats by Position

To find putative introgressed regions we used D statistics (D-stats) with sliding windows (Mondal *et al.*, 2016). We calculated D-stats for 50 kb regions with 5 kb of offset. We specifically looked for regions where Australians lacks African Alleles compared to Europeans:

1. D(AUS,FRN;MAD,Ancestral).
2. D(AUS,FRN;MBT,Ancestral).
3. D(AUS,FRN;SAN,Ancestral).
4. D(AUS,FRN;YRI,Ancestral).
5. D(AUS,DAI;MAD,Ancestral).
6. D(AUS,DAI;MBT,Ancestral).
7. D(AUS,DAI;SAN,Ancestral).
8. D(AUS,DAI;YRI,Ancestral).

To mask for Denisova introgressed regions we used:

9. D(AUS,DAI;DENI,Ancestral).
10. D(AUS,RFRN;DENI,Ancestral).

We only kept regions which gives complete -1 for these D-stats values. To remove low coverage regions, we removed any region which had less than 50 SNPs. For the ancient Andamanese individual, we used the same concept of a dearth of African ancestry compared to Europeans although we did not use masking for Denisova region.

S* test

After getting putative regions for every individual, we used S* to further refine the regions (Vernot et al., 2014). We first calculated a null distribution of no introgression from a known demographic model (Gravel et al., 2011) and replacing East Asians with Aboriginal Australians. As it is impossible to do simulation for every region (having different number of segregating alleles and different recombination rate), we used a generalized linear model to predict S* values for any such arbitrary region (Vernot et al., 2014). We simulated from 11 to 414 segregating sites with a step of 13 and recombination rate was simulated from 0.000155 to 12.900155 with steps of .43. We used Yoruba as a reference population and used one Aboriginal Australian at a time to detect the introgressed regions. We used any region which have higher S* value than the 95 percentile of the null distribution model of that region.

Building a Reference Genome

After getting introgressed regions from both D-stats and S*, we tried to rebuild the reference genome of this unknown hominin population for Aboriginal Australians. We merged all the positions which gave statistical significance for all the individuals. In case of multiple individuals showed positive for same regions we took the highest S* value for that regions and kept those positions for the individuals which gave highest S* value. After getting all the positions together (after removing all the transitions) we used SNPRelate (Zheng et al., 2012) to plot a simple DNA distance matrix tree. As for Andamanese we have only one individual, we did not have to rebuild the reference genome by merging multiple individuals.

References

Abecasis,G.R. *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **135**, 0–9.

Barker,G. *et al.* (2007) The ‘human revolution’ in lowland tropical Southeast Asia: the antiquity and behavior of anatomically modern humans at Niah Cave (Sarawak, Borneo). *J. Hum. Evol.*, **52**, 243–261.

Clarkson,C. *et al.* (2015) The archaeology, chronology and stratigraphy of Madjedbebe (Malakunanja II): A site in northern Australia with early occupation. *J. Hum. Evol.*, **83**, 46–64.

Gravel,S. *et al.* (2011) Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 11983–11988.

Green,R.E. *et al.* (2010) A draft sequence of the Neandertal genome. *Science*, **328**, 710–722.

Malaspinas,A.-S. *et al.* (2016) The genomic history of Indigenous Australia. *Nature*, **160100007**, 1–152.

Mallick,S. *et al.* (2016) The Simons Genome Diversity Project : 300 genomes from 142 diverse populations. *Nature*.

Matsumura,H. and Oxenham,M.F. (2014) Demographic transitions and migration in prehistoric East/Southeast Asia through the lens of nonmetric dental traits. *Am. J. Phys. Anthropol.*, **155**, 45–65.

Meyer,M. *et al.* (2012) A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*, **338**, 222–226.

Mondal,M. *et al.* (2016) Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. *Nat. Genet.*

O'Connell,J.. and Allen,J. (2004) Dating the colonization of Sahul (Pleistocene Australia–New Guinea): a review of recent research. *J. Archaeol. Sci.*, **31**, 835–853.

O'Connell,J.F. and Allen,J. (2015) The process, biotic impact, and global implications of the human colonization of Sahul about 47,000 years ago. *J. Archaeol. Sci.*, **56**, 73–84.

Patterson,N. *et al.* (2012) Ancient admixture in human history. *Genetics*, **192**, 1065–1093.

Pickrell,J.K. and Pritchard,J.K. (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.*, **8**, e1002967.

Prüfer,K. *et al.* (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, **505**, 43–49.

Rasmussen,M. *et al.* (2011) An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. *Science*, **334**, 94–98.

Summerhayes,G.R. *et al.* (2010) Human adaptation and plant use in highland New Guinea 49,000 to 44,000 years ago. *Science*, **330**, 78–81.

Vernot,B. *et al.* (2014) Resurrecting surviving Neandertal lineages from modern human genomes. *Science*, **343**, 1017–1021.

Zheng,X. *et al.* (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, **28**, 3326–3328.

Figures

Figure 1: Absence of African allele calculated by D-stats. The mean value of D-stats is signified by square and two standard deviations is signified by the lines. Positive mean X population (names are in the left and geographical position in right) have less African allele (Yoruba) compared to European (French).

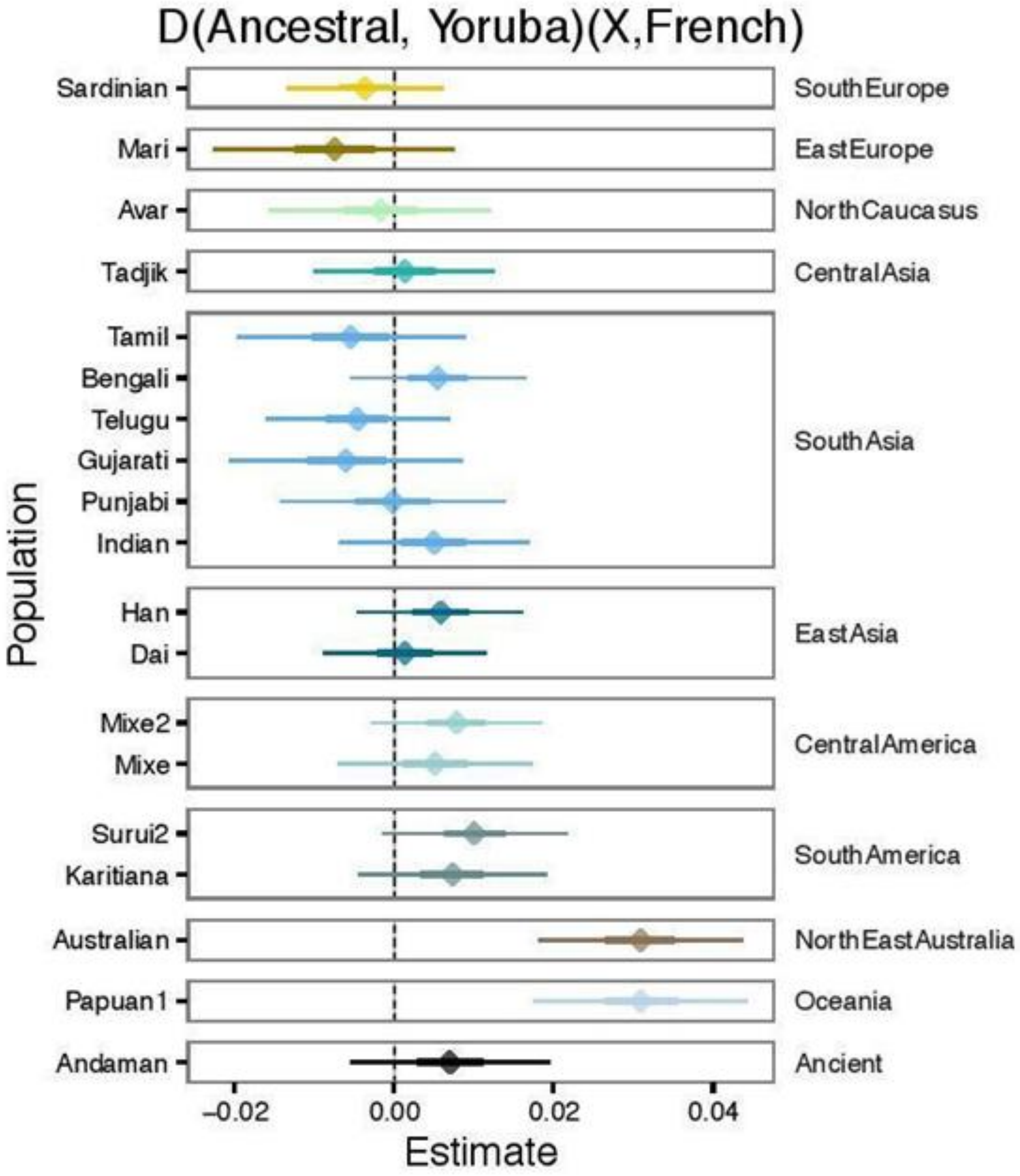
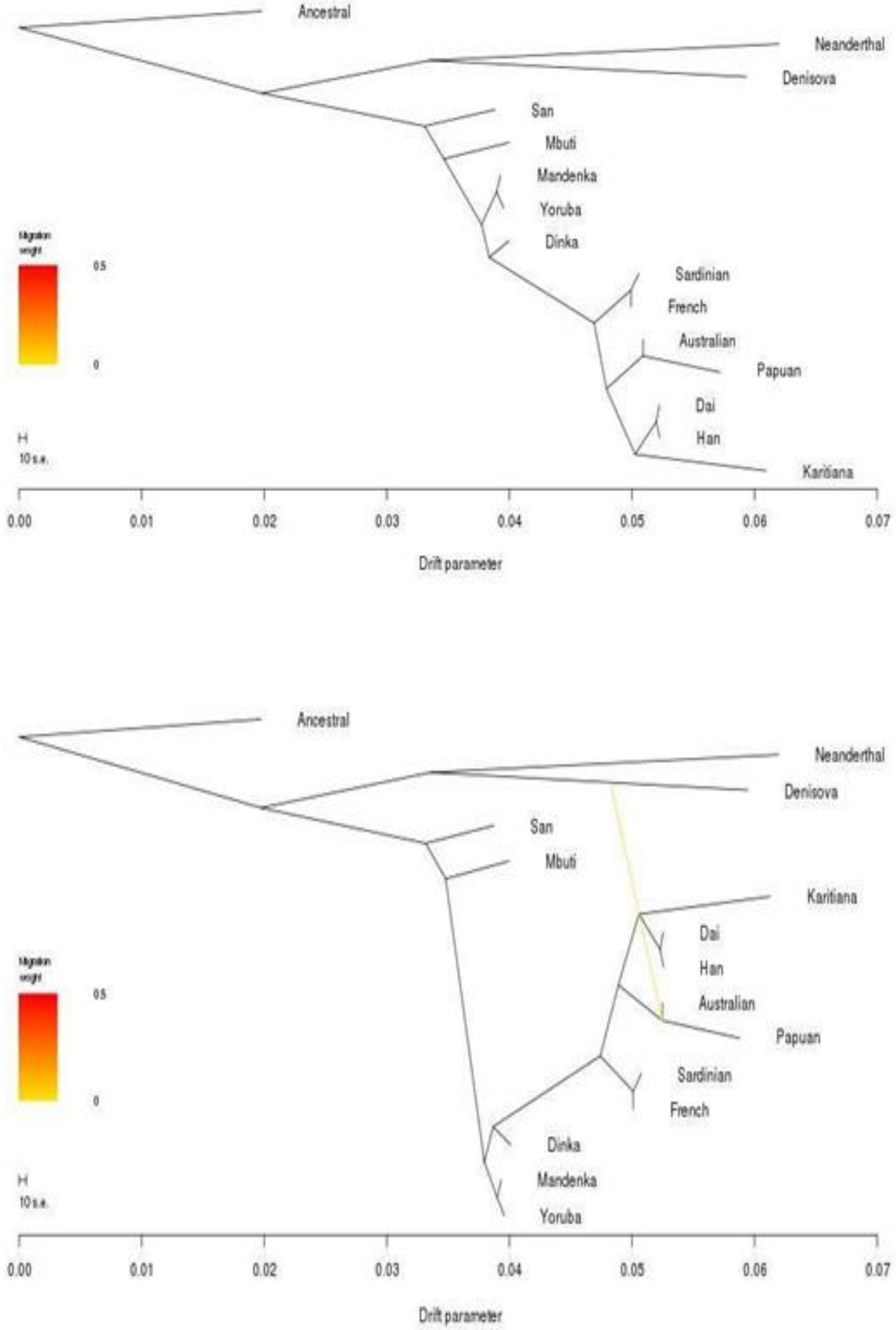
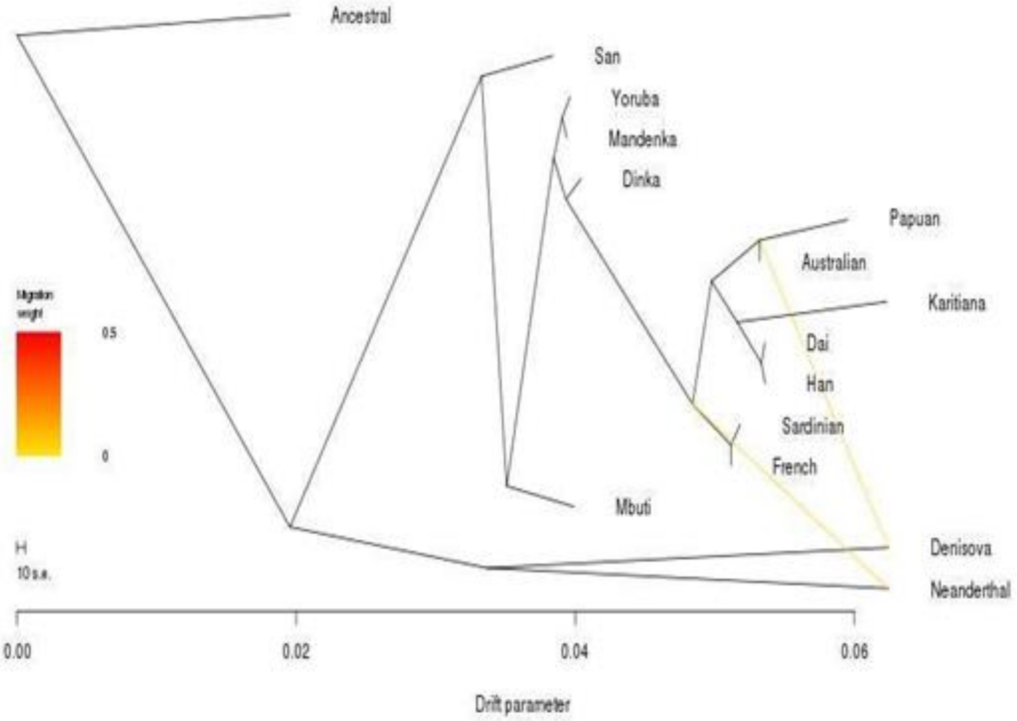


Figure 2: Treemix results using 0-3 migration.




Ancestral
San
Yoruba
Mandenka
Dinka
Papuan
Australian
Karitiana
Dai
Han
Sardinian
French
Mbuti
Denisova
Neanderthal
Migration weight
0.5
0
10 s.e.
0.00
0.02
0.04
0.06
Drift parameter

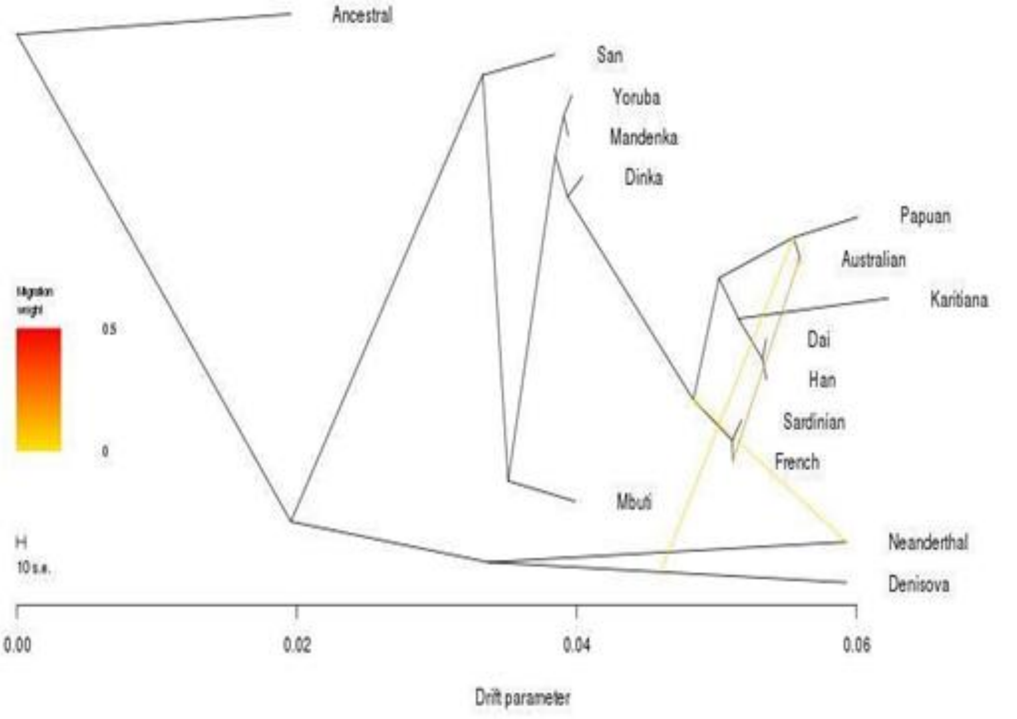

Ancestral
San
Yoruba
Mandenka
Dinka
Papuan
Australian
Karitiana
Dai
Han
Sardinian
French
Mbuti
Neanderthal
Denisova
Migration weight
0.5
0
10 s.e.
0.00
0.02
0.04
0.06
Drift parameter

Figure 3a: Unknown introgressed regions in for Ancient Andamanese.

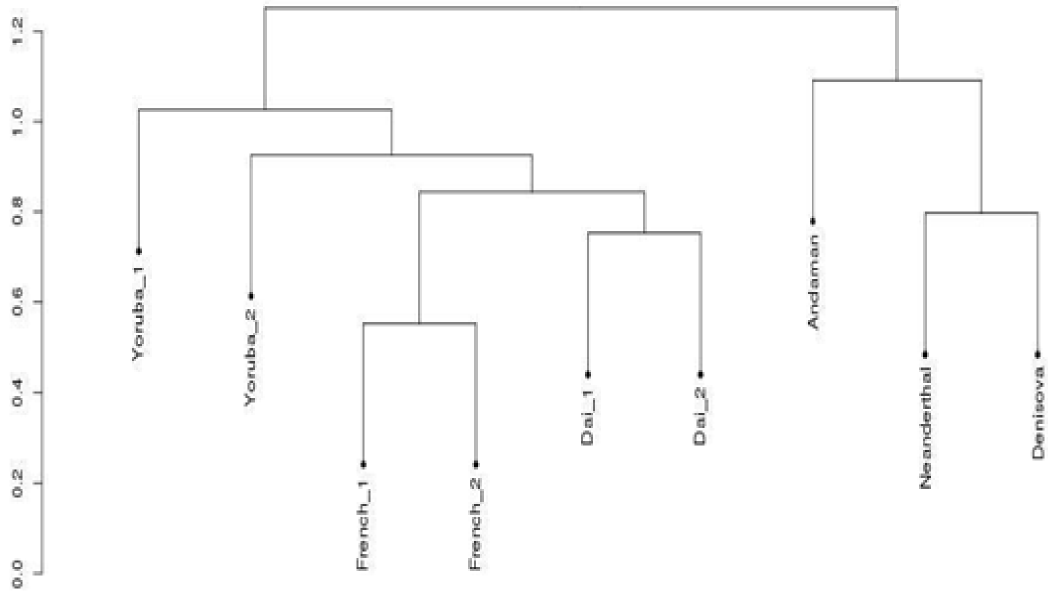
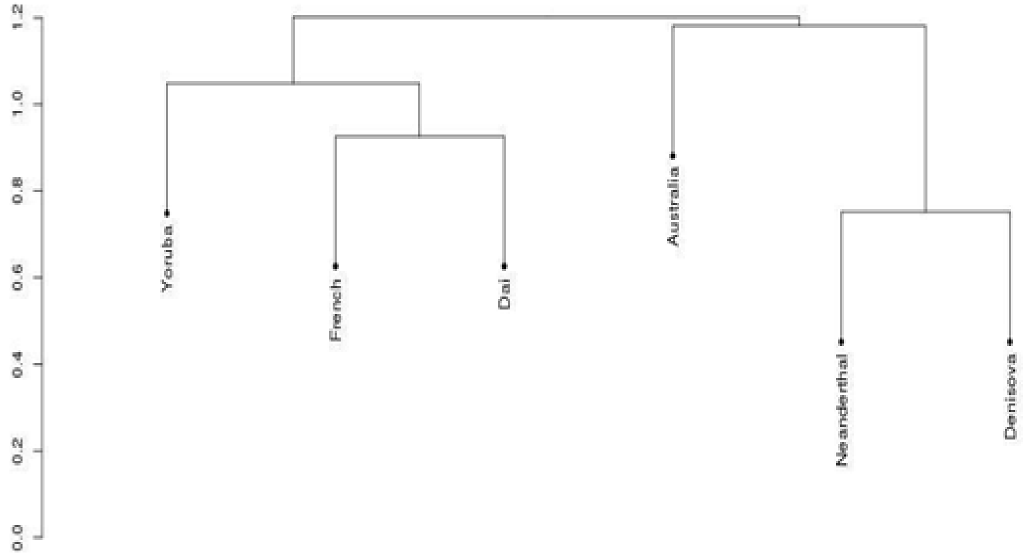


Figure 3b: Unknown introgressed regions in Australia Aboriginals.



Tables

Table 1: Detected Introgressed Regions by different methods for Aboriginal Australians in base pairs

Names	D-stats by region with no masking	D-Stats by region with Denisova Mask	D-stats by region with Denisova mask and S*
BDV01	9,989,849	4,809,932	394,994
BDV02	18,899,720	5,814,894	244,996
BDV04	11,224,849	4,209,934	289,995
BDV05	20,344,694	8,129,857	754,988
BDV06	10,339,846	4,229,931	224,996
BDV07	15,039,788	5,584,904	454,994
BDV08	13,509,797	5,229,908	494,992
BDV09	14,279,789	3,314,939	149,997
BDV10	7,409,878	3,104,944	339,994
CAI01	21,439,715	11,049,846	459,993
CAI02	7,984,866	3,154,946	159,997
CAI03	18,469,743	9,674,852	189,997
CAI04	9,984,850	3,414,945	124,998
CAI05	15,339,784	5,459,913	719,988
CAI06	12,759,812	4,949,919	229,996
CAI07	10,844,819	2,869,945	104,998
CAI08	10,844,837	5,879,910	324,995
CAI09	8,424,862	1,999,965	0
CAI10	12,639,822	4,059,929	279,996
ENY01	4,559,920	2,534,955	124,998
ENY02	17,904,718	8,249,873	514,991
ENY03	14,199,804	9,944,864	509,993
ENY04	12,504,820	5,104,920	244,996
ENY05	10,959,834	7,644,883	744,988
ENY06	13,834,804	8,989,875	799,987
ENY07	11,714,821	4,159,936	149,998
ENY08	10,309,835	4,119,929	339,995
NGA01	13,074,809	3,954,937	314,995
NGA02	14,044,791	5,254,911	444,992
NGA03	10,169,843	4,819,926	279,996
NGA04	14,474,766	7,084,887	854,988
NGA05	12,164,819	5,569,906	369,993
NGA06	8,809,872	2,269,957	69,999
PIL01	16,149,733	4,794,911	364,994
PIL02	12,209,795	4,699,914	259,995
PIL03	14,294,774	5,369,910	179,997
PIL04	9,869,825	4,109,923	209,996
PIL05	17,729,753	6,009,893	329,994
PIL06	10,564,838	3,309,947	184,997
PIL07	16,249,783	5,369,915	514,991
PIL08	16,369,756	6,779,891	419,993
PIL09	11,844,834	2,709,963	74,999
PIL10	13,179,794	4,074,933	269,996
PIL11	9,944,837	2,639,954	109,998
PIL12	9,869,844	3,649,944	289,996
RIV01	17,964,741	8,894,868	629,990
RIV02	11,689,829	6,434,908	144,998
RIV03	12,899,816	7,029,903	314,995
RIV04	10,214,854	5,569,919	659,991
RIV05	13,054,802	4,709,927	129,998
RIV06	10,254,841	5,354,913	484,992
RIV07	14,054,791	5,394,913	379,994
RIV08	10,869,842	7,709,894	634,990
WCD01	15,594,751	4,459,927	199,996
WCD02	14,779,763	5,689,903	169,997
WCD03	18,509,719	5,194,907	344,994
WCD04	17,144,741	2,544,953	199,997
WCD05	19,559,737	5,029,910	299,995
WCD06	25,009,631	5,754,901	339,994

WCD07	19,489,703	6,044,900	434,993
WCD08	18,319,721	5,504,905	319,995
WCD09	18,989,716	7,229,893	624,992
WCD10	14,509,770	3,394,944	164,997
WCD11	15,054,750	5,654,905	289,995
WCD12	18,054,739	6,629,883	239,996
WCD13	17,444,734	5,259,908	244,996
WON01	19,404,745	4,499,938	339,995
WON02	9,039,863	5,099,924	419,994
WON03	10,604,825	5,549,910	214,996
WON04	13,869,812	5,794,915	459,993
WON05	15,914,777	5,419,904	249,996
WON06	19,549,722	6,019,895	389,993
WON07	10,219,860	3,564,940	264,996
WON08	13,419,799	8,229,872	649,989
WON09	11,129,852	2,979,958	169,998
WON10	10,624,855	3,734,943	354,994
WON11	6,089,902	3,739,945	159,998
WPA01	19,559,724	8,924,862	534,991
WPA02	16,534,745	4,549,922	479,992
WPA03	6,094,896	2,549,954	104,998
WPA04	15,579,797	7,989,877	784,989
WPA05	14,279,772	4,364,921	204,997
WPA06	5,039,911	1,399,974	99,998
Total	138,298,233	76,149,009	16,694,760
Andamanese	37,214,558	NA	2,579,962

3.6 Y Chromosome profile of Indian continental populations and ancestry dilemma of Andamanese Populations

Mayukh Mondal et al.

Manuscript in Preparation

Introduction

Y chromosome is a powerful tool to analyse the paternal ancestry of human populations. As most of the Y chromosome does not recombine; reconstructing haplotypes are much easier. It is possible to reconstruct, in a deterministic way, the gene tree for all human variation that analysed in a geographic context. This allows a phylogeographic approach which now may encompass not only the specific variants that defined the classical “haplogroups”, but that contains all the nucleotide variation in the whole chromosome.

Indian continental population ancestry is complex and many attempts have been done using both uniparental markers (mtDNA and Y-chromosome) and autosomes. The Y-chromosome analysis has mostly relied on calculating frequencies of pre-defined haplogroups. The time and place of origin have been estimated for some of haplogroups. Some haplogroups found in India are rare outside of India, making it difficult for a clear interpretation.

In the present study we have reconstructed the male Indian origin by reconstructing the whole Y chromosome phylogenies using whole genome sequences of a wide set of populations, including new non-tribal and tribal (including Andamanese) populations analysed along the recently produced non-tribal populations of India from 1000 Genome. We also tried to elucidate the apparent stark contrast of Andamanese ancestry (belonging to D haplogroup) in relation to most other Out of Africa (OOA) populations and from autosomal data.

Results

Haplogroups found in India

Indian continental populations showed a complex ancestry from Y chromosome haplogroup analysis (Figure 1 and Table 2). Unlike Europeans (where main haplogroup is R), East Asians (where main haplogroup is O) or Africans (where main haplogroup is E), Indian populations have no single major haplogroup. Major haplogroups in Indian continent include R, C, H, J and L. Other haplogroups found in the Indian continent are D, G, K, L, N, O, Q. These haplogroups do not share a common simple ancestry (Figure 3), suggesting a complex origin of the many populations in India, with different founder populations and admixture.

Main Haplogroups

Haplogroup R is present in high frequency in all non-tribal populations (both Indo-European and Dravidian speaking populations) but this haplogroup absence in all tribal populations suggests that haplogroup R did not arise within India but rather was brought to India from outside, as was suggested before (Zhao et al., 2009). One of the main sources of this haplogroup could be Indo-European (Aryan) migration around five kilo years ago (kya) (Thapar, 1996). Nonetheless it is interesting to note that there is not a strong North-South cline in the frequency of R as would be expected under the hypothesis of a north migration: the frequency expected in the South should be lower than that observed in 1000Genomes populations STU and ITU.

A better insight may come from considering the sub-haplogroups of R. Indeed, when calculating the minimum divergence time of Indian Haplogroup R1 with non-Indian (presumably Europeans), we obtained a value of around eight kya. Nonetheless, haplogroup

R2, which is only found in India and not found in European populations from 1000 Genomes, have divergence within India is around 10 kya. Thus the entry in India must be much earlier than the entrance of Indo-Europeans, with high and similar frequencies in non-tribal populations from many different places in the wide Indian geography.

Other main haplogroups which are found in India is C, H, J and L. Haplogroup C is found mostly in Gujaratis (GIH) with lower frequencies in other northern populations. H also presents a frequency with small North-South differentiation and presents both in tribal and non-tribal populations. J again with no much geographic stratification, is found in non-tribal populations. And L haplogroup is mainly found in the South, in tribal and non-tribal populations. In our tribal populations (Irula [ILA] and Birhor [BIR]) both H and L are found suggesting a common Indian ancestry at least for these two haplogroups.

Minor Haplogroups

D haplogroup is exclusive for Andamanese and will be discussed later. G haplogroup is found at low frequency in the Northwest (PJL and GIH). Both K and N haplogroups are only found at very low frequency in ITU. Haplogroup O is very interesting, as it accounts for all the chromosomes in the Tibeto-Burman population (RIA) and found in 10% of Bengalis from Bangladesh (BEB), being thus a haplogroup of the East of the region. This distribution and the frequency in Eastern populations suggest a recent migration from East Asia that entered India via the North Eastern border; the minimum time divergence of these haplogroup with Non-Indian populations is around eight kya. Finally, haplogroup Q is found in very low frequency in most Indian populations.

Andamanese

One of the most interesting haplogroups found in Indian tribal populations is Haplogroup D, which is found only in Andamanese in India and in all five individuals sequenced. This haplogroup is especially interesting as it has an around 4000 years more recent ancestry with African E haplogroup compared to all other haplogroups found in OOA populations (Poznik et al.). This recent ancestry with African populations has been a base for postulating a first OOA migration, differentiated from the later and more widespread East Asian expansion (Thangaraj et al., 2003; Shi et al., 2008). This result is stark contrast to what we expect from Autosomal data where Andamanese showed a more recent ancestry with Asian populations without a trace of contribution from a putative first OOA (Mondal et al., 2016). This haplogroup is also found at high frequencies in Japanese in Tokyo (JPT) (present in the 1000Genomes data) and in Tibet (without sequence data). Time divergence between Andamanese and JPT individuals having D haplogroup (JPTD) is ~ 54 kya and time divergence between Haplogroup D and O (Japanese having O haplogroup or JPTO) is ~ 76 kya (Figure 4). This discrepancy between Y chromosome and autosomal data can be explained by two different hypotheses: 1) it can be caused by two out of Africa events. People having haplogroup D first populated the Andaman islands and Japan and later in Japan it was replaced by Haplogroup O individuals (JPTO) which would have a different out of Africa origin; haplogroup D is more frequent in the two extremes of the archipelago as a consequence of a later substitution(Hammer et al., 2006) and 2) when a unique OOA event happened haplogroup D along with other haplogroups (which are found in other OOA populations, i.e. C and F) was already present in OOA populations but later Haplogroup D was removed from other OOA populations by random chances and left only with Andamanese and Japanese individuals(Poznik et al., 2016). If the first hypothesis is true we would expect that Andamanese and Japanese having D haplogroup would share more derived alleles with each other than to other East Asian populations (for example Dai from China, or even with JPTO individuals) in the autosomes. We found that is not the case (Table 3). In fact, Andamanese is an out group of Asian populations (at least for DAI and Japanese). We also build a simple simulation model from Y chromosome data, where Andamanese and JPTD having separated from other OOA populations around 76 kya and separated between themselves around 54 kya. Later around 11 kya (400 generation) JPTO individuals started to admix with JPTD individuals without affecting Andamanese. To get the result of empirical D-stats we need 99 % (±3%) admixing proportion from JPTO populations to JPTD, which is very big and unlikely, leaving us with the 2nd hypothesis. The high frequency of D haplogroup in Tibet (Gayden et al., 2007 and references therein) seem to be part of a different expansion, even if new studies with sequence data are needed to have a complete picture.

Conclusions

We were able to show that Indian populations showed a very complex ancestry which cannot be explained by a single expansion event populating whole Indian continent; even that, Y-chromosome data shows less diversification between the North and South than what has been described in autosomal studies. There is a recent ancestry (~8kya) shared with Europeans and Indians via haplogroup R1 which is likely related to the Indo-European (Aryan) migration. Indian tribal populations have a complex ancestry having shared C, H, J and L haplotypes. As these haplogroups were separated from each other not less than ~50 kya, it is more probable that when Indian continent was populated these variations were already present in the populations suggesting that the whole Indian subcontinent was populated not before 50 kya. The alternative hypothesis would be that India was populated several times independently with populations having high frequency of C, H, J and L haplogroups, which would be more difficult to imagine though not impossible. Haplogroup O in the Northeast (BEB) is a recent introduction (around eight kya) most probably by a Tibeto Burman population.

Andamanese and some Japanese (but also Tibetans) showed an interesting haplogroup D, which was one of the main reasons of genetic study to postulate Andamanese as an out-group of all out of African populations following a coastal route migration from which only extreme populations would have subsisted. We were able to show that Andamanese and Japanese individuals having D haplogroup have separated around ~ 54 kya coinciding with most of the other major Out of Africa haplogroup divergence. We also showed, using autosomal data, that Andamanese is indeed an out group of East Asian populations, which strongly suggest Haplogroup D does not show a real separate ancestry for Andamanese populations. Rather it is a part of the standing variation when out of Africa event happened and later removed from most of the populations except in Andaman and, partially, Japan and Tibet.

Methods

Samples

In total 42 samples of 10 different Indian populations (including 5 Andamanese) were used in the analysis (Table 1 and Figure 1). For more information about the populations, see Mondal et al (2016). We also used 1000 Genome populations to compare with our data (Poznik et al. 2016).

Sequencing

The whole genome sequencing of Indian populations was done by two different institutes using Illumina technology (Mondal et al., 2016). We extracted Y chromosome sequences from the bam file of whole genome sequences using samtools 1.1 (Li et al., 2009). The average coverage of Y chromosome for Indian data is close to ~15x per individual. Although for the regions, which are suitable for short read sequencing (Poznik et al., 2013), the average coverage is ~7x (Figure 2). 1000 Genome chromosome Y bam files of 1244 individuals from phase 3 was download from 1000 Genome project site (Auton et al., 2015).

Variant Calling

Variant calling on the bam files of India and 1000 Genomes was done using Genome Analysis Toolkit 3.5 (McKenna et al., 2010) using HaplotypeCaller and gvcf method using default parameters except ploidy of the genome was set as 1. After getting individuals gvcf files for all the individuals we called them together by GenotypeGVCFs. We used dbsnp version 137 to get the rsid for known SNPs (Sherry et al., 2001). Calling was restricted to the regions which are suitable for calling for Y chromosome (Poznik et al., 2013). We did the variant calling for all the sites using -allSites flag in the GenotypeGVCFs. Other parameters were set to default.

Filtering

As 1000 Genome has low coverage, we put several filters to get the positions where we have good power to do the variant calling. We removed any position which has lower coverage in total than 1502x (half of average coverage for all the sites) and more than 6006x (double of the average coverage for all the sites). If a polymorphic site has more than 200 individuals having a different read than the called genotype, it was also removed. If a position had a greater than ratio of 0.1 for the number of reads with mapping quality 0 to the total number of reads it was also removed. Any position having more than 30% of sample with missing genotype was also removed.

Phylogenetic inference and dating

After using the above filters and removing all the indels, we removed all the monomorphic sites from the data set. We are left with 61,924 sites for the whole data set. We then used

PGDSpider-2.0.9.2 (Lischer and Excoffier, 2012) to convert vcf file to phylip format which is required for RAXML input. We used RAXML 8.2.4 (Stamatakis, 2014) for the phylogenetic analysis. We used ASC_GTRGAMMA model of nucleotide substitution with "stamataki" correction for ascertainment bias. We used HG02982 and HG01890 as an out-group of all modern humans, which was already shown to have A0 haplogroup (Poznik et al., 2016). We used 100 bootstrap replicates to calculate statistical support and visualized it with Tree of Life (Letunic and Bork, 2011).

In RAXML output distance were measured with the number of mutations, which were transformed to number of years to know the time of divergence between haplogroups. We calculated the average distance with A0 individuals with all other individuals using cophenetic.phylo from "ape" R package (Paradis et al., 2004). As we know the time of divergence of A0 from other individuals (Poznik et al., 2016), we just multiplied all the distances with a suitable number to convert genetic distance with the time of divergence for every individual.

Pie chart analysis

We used "maps", "mapdata", "mapproj" from R package to plot Indian continent maps and "plotrix" and "Rcolorbrewer" to plot the pie charts.

Whole Genome Sequence analysis

The 10 Japanese bam files (5 having haplogroup D and 5 having haplogroup O) for autosomes were downloaded from the 1000 Genome site. Andamanese individuals and DAI BAM files were accessed from the previous project (Mondal et al., 2016). The variant calling was done in similar way that we have done on chromosome Y. We changed the ploidy option 2 for HaplotypeCaller and do the variant calling for only the polymorphic SNPs in our data set. We used the VariantRecalibrator from GATK using dbsnp137, HapMap 3.3, 1000 Genomes Project Omni 2.5 and 1000 Genomes Project Phase 1 SNPs with high confidence downloaded from the Broad Institute ftp site (ftp.broadinstitute.org, 11/05/2013). After VariantRecalibrator was done we only kept the SNPs which has passed the filter and only kept SNPs which has no missing information. We added ancestral information from 1000 Genome Project website (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/). We then used ADMIXTOOLS 1.1 (Patterson et al., 2012) to calculate D-stats for autosomal data.

Simulation

We build a simple model where JPTO have separated from AND and JPTD 76 kya ago and AND and JPTD separated from each other 53 kya following chromosome Y analysis. We used a mutation rate (μ) of 1.25×10^{-8} per site per generation, recombination rate (r) of 1.3×10^{-8} per site per generation and generation time of 29 years. As D-stats is neither affected by the effective populations size nor by time of admixture (Patterson et al., 2012), we put effective populations size of all these populations to be 10,000 and time of admixture from JPTO to JPTD around 400 generations ago. We simulated 30,000 regions of 50 kb using ms (Hudson, 2002):

```
ms 40 30000 -t 50 -r 52 100000 -I 3 20 10 10 -es 0.0175 2 <VAR> -ej 0.0175 4 3 -ej 0.0457
2 1 -ej 0.0655 3 1
```

Where <VAR>=0-.99 with step of .01.

The D-stats values were calculated from the simulated data. The fitting of the data was done by "lm" from the R package.

References

Auton,A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Gayden,T. *et al.* (2007) The Himalayas as a directional barrier to gene flow. *Am. J. Hum. Genet.*, **80**, 884–94.

Hammer,M.F. *et al.* (2006) Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. *J. Hum. Genet.*, **51**, 47–58.

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

Letunic,I. and Bork,P. (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.*, **39**, W475-8.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lischer,H.E.L. and Excoffier,L. (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, **28**, 298–9.

McKenna,A. *et al.* (2010) The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

Mondal,M. *et al.* (2016) Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. *Nat. Genet.*, 1–102.

Paradis,E. *et al.* (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, **20**, 289–290.

Patterson,N. *et al.* (2012) Ancient admixture in human history. *Genetics*, **192**, 1065–1093.

Poznik,G.D. *et al.* (2016) Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat. Genet.*, **12**, 809–809.

Poznik,G.D. *et al.* Punctuated bursts in human male demography inferred from 1 , 244 worldwide Y-chromosome sequences Main Text. 1–16.

Poznik,G.D. *et al.* (2013) Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science*, **341**, 562–5.

Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

Shi,H. *et al.* (2008) Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. *BMC Biol.*, **6**, 45.

Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–3.

Thangaraj,K. *et al.* (2003) Genetic Affinities of the Andaman Islanders, a Vanishing Human Population. *Curr. Biol.*, **13**, 86–93.

Thapar,R. (1996) The Theory of Aryan Race and India: History and Politics. *Soc. Sci.*, **24**, 3–29.

Zhao,Z. *et al.* (2009) Presence of three different paternal lineages among North Indians: a study of 560 Y chromosomes. *Ann. Hum. Biol.*, **36**, 46–59.

Figures

Figure 1. Approximate positions of the studied Indian populations with a haplogroup composition pie chart. Size of the pies: small, one individual; intermediate, 4-8 individuals; big, 42-60 individuals from 1000G

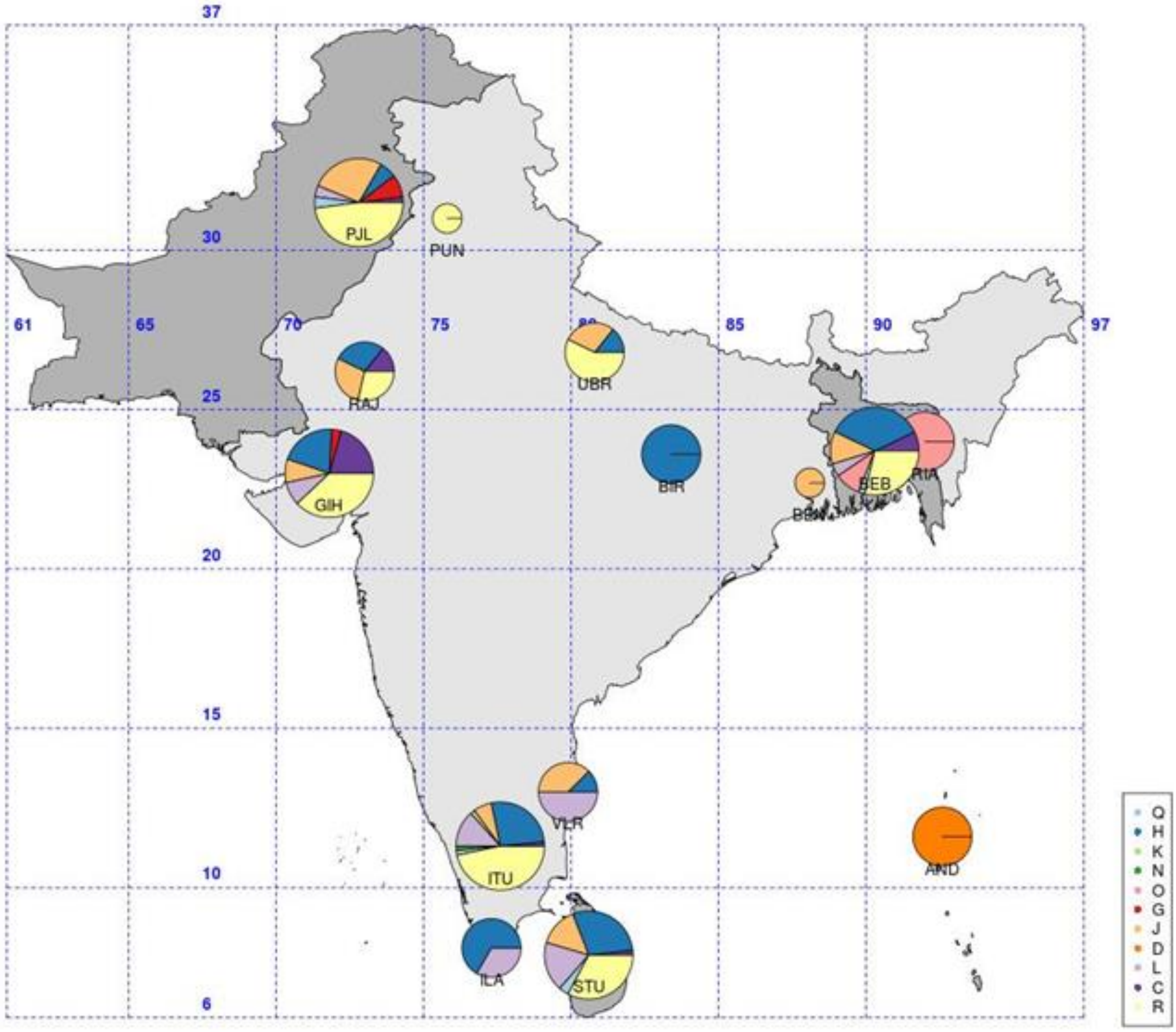


Figure 2: Coverage of the 42 Indian samples

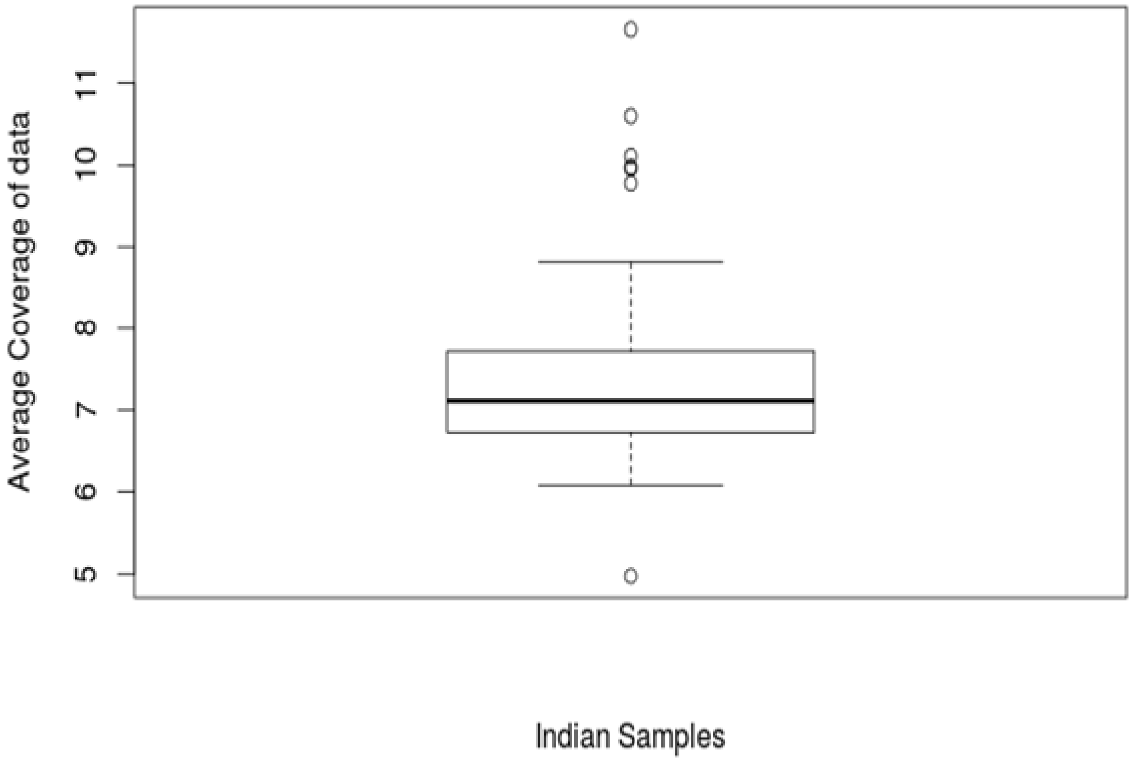
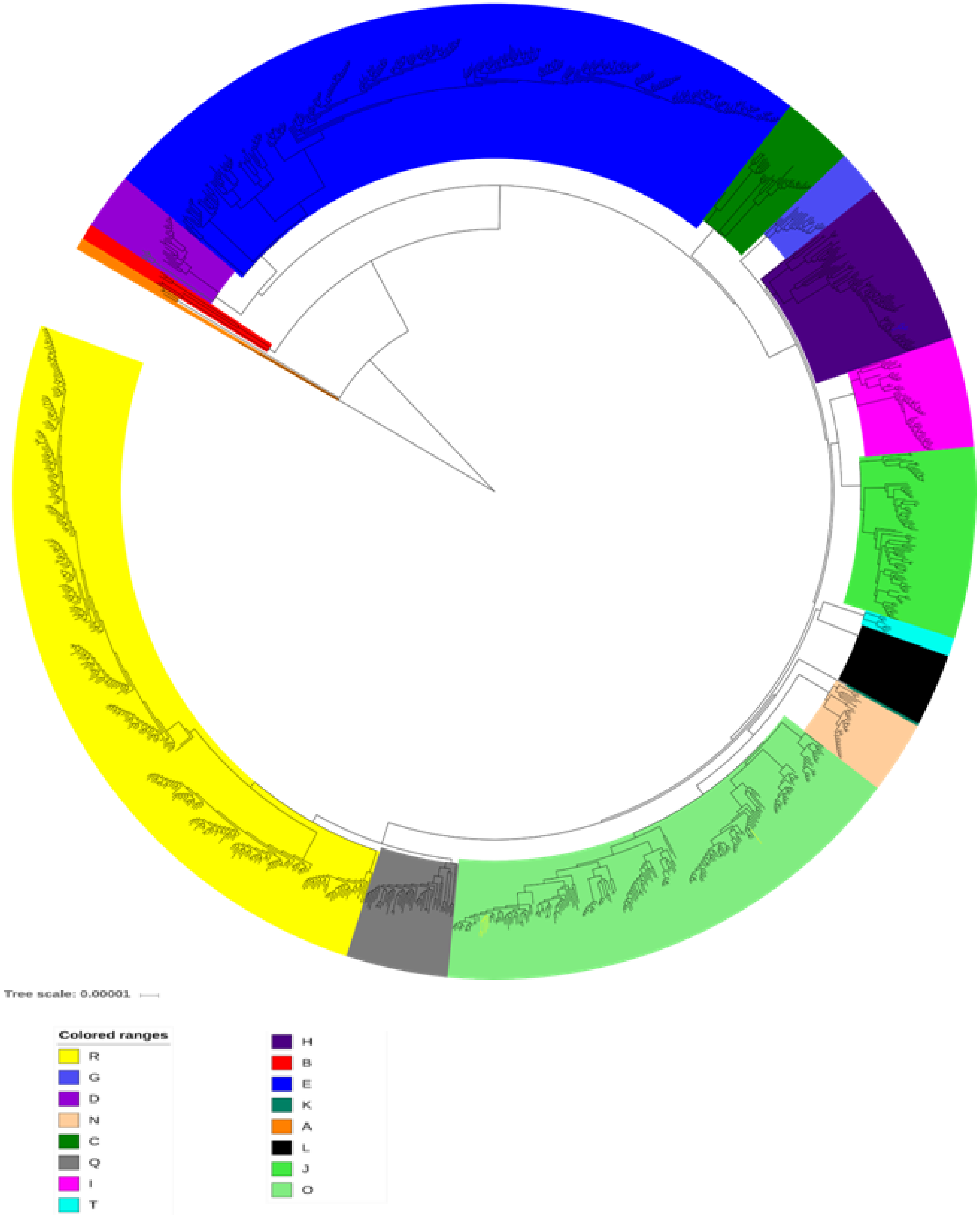
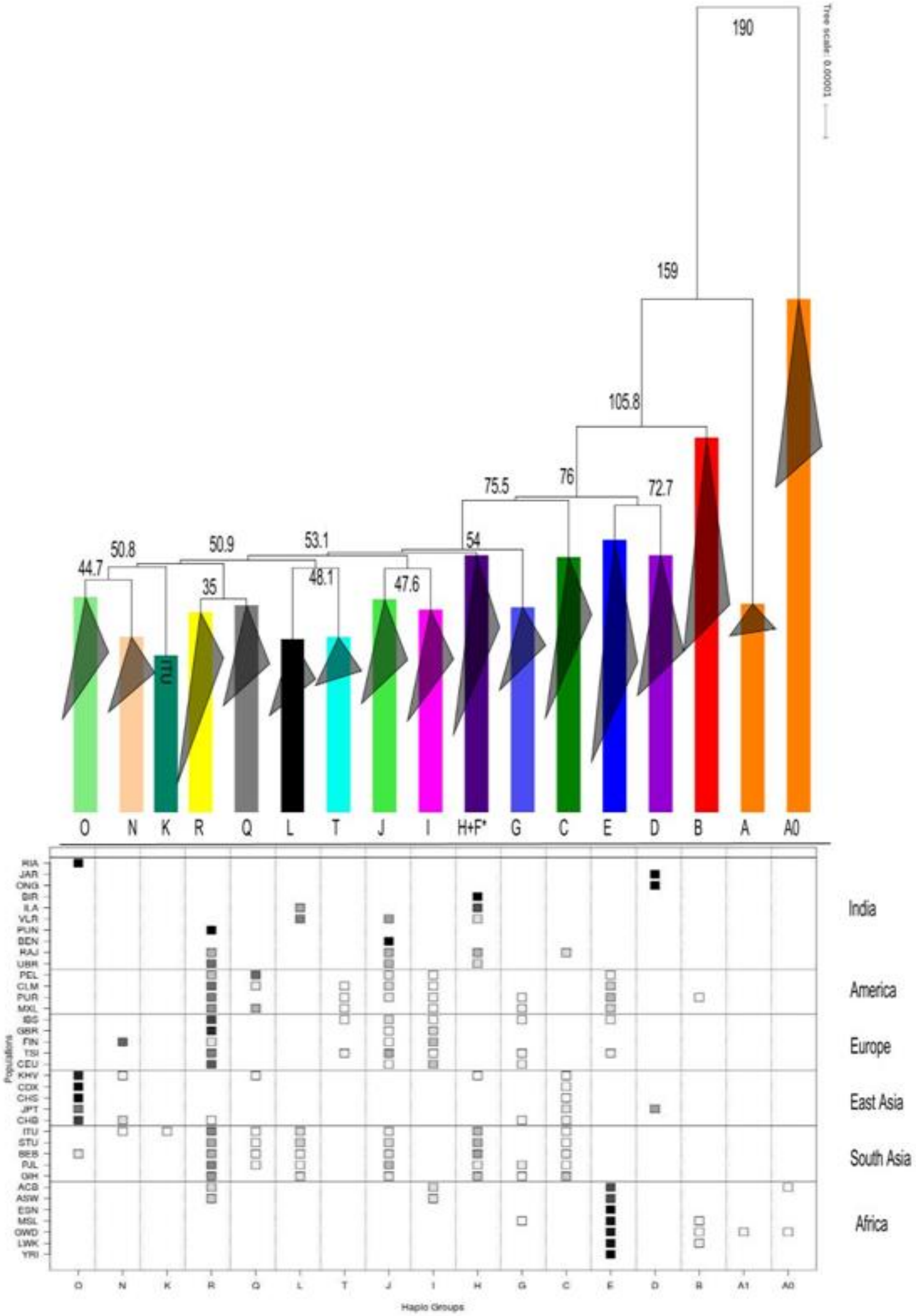


Figure 3: Maximum likelihood phylogeny for 1000 Genome and Indian data (a) all individuals, (b) collapsing major branches in haplogroups

(a)

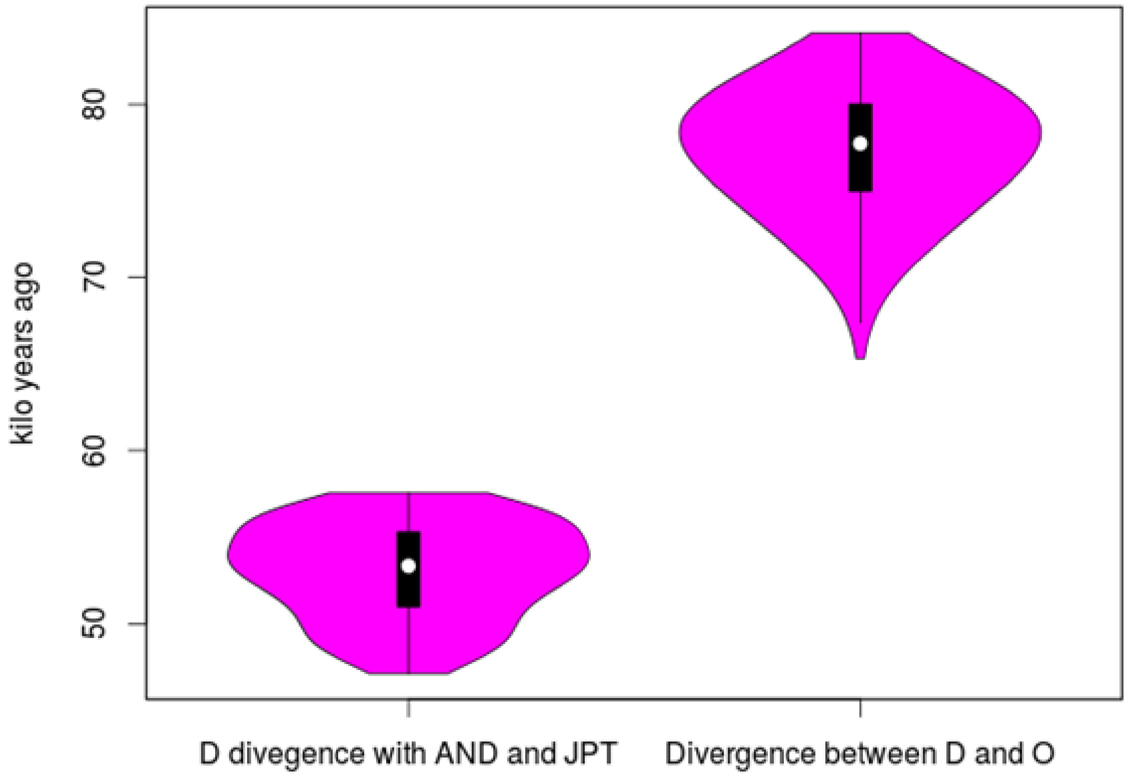


(b)



Tree scale: 0.00001
190
159
105.8
75.5
76
72.7
53.1
54
50.9
50.8
44.7
35
48.1
47.6
ITU
O
N
K
R
Q
L
T
J
I
H+F*
G
C
E
D
B
A
A0
RIA
JAR
ONG
BIR
ILA
VLR
PUN
BEN
RAJ
UBR
PEL
CLM
PUR
MXL
IBS
GBR
FIN
TSI
CEU
KHV
CDX
CHS
JPT
CHB
ITU
STU
BEB
PJL
GIH
ACB
ASW
ESN
MSL
GWD
LWK
YRI
Populations
India
America
Europe
East Asia
South Asia
Africa
O
N
K
R
Q
L
T
J
I
H
G
C
E
D
B
A1
A0
Haplo Groups

Figure 4: Divergence time between Y-chromosomes belonging to the D haplogroup, one from Andaman and the other from Japan (left) and between chromosomes one from haplogroup D and the other O, both from Japan.



Tables

Table 1: Sample size information of Indian populations

Population	**Geographical Region**	**Linguistic Affiliation**	**Social Category**	**No. of Individuals** **Sequenced**
Brahmin (UBR)	North	Indo-European	Upper Caste	7
Rajput (RAJ)	North	Indo-European	Middle Caste	7
Bengali (BEN)	North	Indo-European	Lower Caste	1
Punjabi (PUN)	North	Indo-European	Middle Caste	1
Vellalar (VLR)	South	Dravidian	Middle Caste	8
Irula (ILA)	South	Dravidian	Tribe	3
Birhor (BIR)	Central	Austro-Asiatic	Tribe	4
Onge (ONG)	Andaman & Nicobar Islands	Unclassified	Tribe	2
Jarawa (JAR)	Andaman & Nicobar Islands	Unclassified	Tribe	3
Riang (RIA)	North-east	Tibeto-Burman	Tribe	6

Table 2: Haplogroup composition of Indian populations

Population	C	D	G	H	J	K	L	N	O	Q	R
AND	0	1	0	0	0	0	0	0	0	0	0
BIR	0	0	0	1	0	0	0	0	0	0	0
ILA	0	0	0	0.67	0	0	0.33	0	0	0	0
VLR	0	0	0	0.12	0.38	0	0.5	0	0	0	0
RAJ	0.14	0	0	0.29	0.29	0	0	0	0	0	0.29
BEN	0	0	0	0	1	0	0	0	0	0	0
UBR	0	0	0	0.14	0.29	0	0	0	0	0	0.57
RIA	0	0	0	0	0	0	0	0	1	0	0
PUN	0	0	0	0	0	0	0	0	0	0	1
PJL	0.02	0	0.08	0.06	0.27	0	0.04	0	0	0.04	0.48
GIH	0.21	0	0.03	0.21	0.09	0	0.09	0	0	0	0.38
BEB	0.07	0	0	0.36	0.12	0	0.05	0	0.1	0.02	0.29
STU	0.02	0	0	0.29	0.15	0	0.18	0	0	0.04	0.33
ITU	0.02	0	0	0.27	0.07	0.02	0.13	0.02	0	0.02	0.47

Table 3: D-stats of Andamanese, DAI, JPTD, JPTO and YRI

W	X	Y	Z	D score	Z score
DAI	JPTD	AND	YRI	0.0028	0.718
DAI	JPTO	AND	YRI	0.0059	1.521
JPTO	JPTD	AND	YRI	-0.0033	-1.475
DAI	JPTD	AND	Ancestral	0.0043	1.214
DAI	JPTO	AND	Ancestral	0.0059	1.647
JPTO	JPTD	AND	Ancestral	-0.0017	-0.856

4. Discussion

After 4 years of constant work, my thesis has come to an end (we still have some loose ends, which I will try to solve in the future). As after reading the whole thesis, it is clear that my work mainly concentrated on bioinformatic works. In this thesis, I have never done substantial wet lab work (except putting some vials from here and there). I guess a few years ago this kind of PhD on Biomedicine would be completely unheard of. But after the publication of Human genome (Lander *et al.*, 2001; Venter *et al.*, 2001), cascades of changes are happening in biology for last 15 years. Now we can able to do PhD in biology without touching any biological stuff. My hope this is the beginning of theoretical biology gaining its power, which it deserved (specifically evolution, which is always less appreciated than other more practical branches of biology). It would transform biology completely and would elevate it from descriptive science to conceptual science.

Biology is famous for unknown parameters. It is not like Newtonian mechanics where few object or parameters can explain most of the things. However, when you have 3 billion data points of information per individual (i.e. a whole genome sequence), we can handle unknown parameters and make a generalised model for human populations with the help of proper statistics. In this thesis, all of the analyses are simply a testament of that. We create general models of population's history and adaptation. Of course, the reality would be much more complicated but these generalisations can give us some insight into how we evolved as modern humans (especially Indian populations for this thesis).

4.1 Digging deeper inside Ancestry (Aryan vs Dravidian), Portability and Consanguinity of Indian population

Availability of sequencing data was scarce for Indian populations when we published our first paper. We are one of the first few to report some population genetic study using exome data on the Indian population. As this being our pilot project, we attempted to test the power of exome sequencing by comparing with genotyping data. We discovered high concordance level (~99%) between genotyping and sequencing data (at least for this study). We also performed a genetic marker portability test of Indian populations. This has high importance for genotyping data (not so much important for sequencing data, as they do not have ascertainment bias). We exhibited that the portability is low for Indian population from European, East Asian or African populations. However, between North and South India, we found high similarity suggesting they are portable between each other. This is a stark contrast to what it was believed earlier. As Indian populations, generally show high Fst value within themselves it was thought that they would not be compatible with each other and thus would fail the portability test. This has importance for medical genetic studies in India, which is still in infancy. This exhibited that although Indian populations do have substructure because of admixing between Aryan and Dravidian components (which we also independently conveyed in the paper), essentially, they are much closer to each other than other continental reference populations. We also found higher inbreeding coefficient in Indian populations compared to European population (CEU). This can be caused by the caste system and/or consanguineous marriages, which is a custom in some part of India. In some part of South India, marriage between Uncle and Niece takes place. This might explain higher inbreeding coefficient detected in South India compared to North India. Indeed, we found a couple from South Indian population to be 12.5% related to each other. This consanguinity results are directly obtained from their genome rather than using genealogical studies. Thus, it is less biased and gives an idea of the assortative mating pattern in Indian populations. We also discovered few interesting genes under selection. The most prominent one is SLC24A5 which is related with skin pigmentation and already known to be one of the main component to give light skin pigmentation in Europeans (Lamason *et al.*, 2005). We found the frequency is high for the derived allele of having light skin pigmentation in North India (~90%). It might be debatable the origin of such selective sweep. This variant might be selected inside India or rather reached India because of a selection process happened earlier in Europe and then brought to India by Aryan migration. It might also be caused by sexual selection still ongoing in India. As light skin pigmentation generally regarded as a desirable characteristic in the spouse.

4.2 Understanding Andamanese Ancestry and Adaptation due to Insularity

This is the main paper of my thesis. Earlier we thought that we would publish a single paper explaining Indian populations' history using genetic materials but after starting the work we realised Andamanese are not directly related to mainland Indian populations. We also found an unknown hominin introgression in Andamanese, which deserved a paper on its own. Therefore, we decided only to concentrate on Andamanese for this paper. We made three very important discoveries (all of which single-handedly have merit to be published as a single paper):

4.2.1 No First OOA Remnant in Andamanese

Right now, the scientific community has a raging debate about the concept of OOA migration for modern humans. Some school of thought believe that two OOA event has happened for modern humans. This hypothesis was mainly supported by anthropological studies (Armitage *et al.*, 2011; Scerri *et al.*, 2014; Smith *et al.*, 2007) and some genetic studies (Rasmussen *et al.*, 2011; Kuhlwilm *et al.*, 2016). The other school of thought is that there is only one OOA event for all modern humans. Andamanese was thought to be one of the main candidates for first OOA event. It was also already argued that Australian has some admixing with this first OOA population (Rasmussen *et al.*, 2011). We proposed that neither Andamanese nor Aboriginal Australian does have a detectable amount of this first OOA population inside their genome. We exhibited by simulation model that the introgression from hominin population could be detected as a false positive of first OOA admixing with these populations. Of course, there might be a low amount of admixing with Andamanese or Aboriginal Australian populations with the first OOA modern humans but with the currently available methods it is not detectable. We do not argue that the first OOA event did not happen. It is just that these populations do not have any detectable contribution from the first OOA population.

4.2.2 A new Hominin discovered

The main discovery of the paper was to find a new hominin group introgressed in Andamanese, undoubtedly making it the most controversial yet the most interesting discovery of the paper. We developed our own new method with the already known methods to detect this hominin introgression. We here concentrated only on Andamanese which impaired us from finding the full impact of this unknown hominin population. Later we revealed that this hominin has much higher impact on modern humans, rather than only affecting Andamanese (Methods and Results Section 3.4-3.5). We hypothesised this hominin is another unknown hominin out-group of both Neanderthal and Denisova. This is one of the first papers to detect an extinct hominin group without having been sequenced.

4.2.3 Adaptation due to Insularity causing Andamanese having *Negrito* Morphology

After proving that Andamanese did not have different African ancestry compared to other contemporary Asian populations, we focused on their morphology, as Andamanese morphology is distinct from other contemporary Asian populations. We used our in-house Hierarchical Boosting method (Pybus *et al.*, 2015) to detect signatures of Natural selection in Andamanese genome. We found that Andamanese people have gone through strong adaptation with height-related genes. This is interesting as it matches with the hypothesis of "insular dwarfism" hypothesis which generally takes place with the large animal living in an isolated place (Lomolino, 2005). This phenomenon can be explained by different hypotheses. The idea is if a big animal (preferably land-based) starts to live in an isolated place (i.e. island), where the food resource can decline to a borderline level periodically, having a smaller body size would help the population to get past such events due to less consumption of foods. Smaller body size would also help facilitate shorter lifespan and early maturation with relatively low amount of resources. Moreover, as in insular regions chances of having predators are low, growing a smaller body size is an effective strategy to competing for food resources. We do not know what exactly caused the dwarfism in Andamanese. It can be either of these explanations or of combinations of these. It can also happen by other causes, which we have not discovered yet. Nonetheless, we are sure that having a low body size has been advantageous for Andamanese populations.

4.3 Unknown Hominin Revisited and Introgression in East Asians

As our previous paper met with some controversy, we revisited our unknown hominin introgression hypothesis. It looks like one of the tests (absence of African allele by D-stats) is difficult to reproduce and there were also some results we could not explain at that time (like how seemingly unrelated populations from South [i.e. Indian] and Southeast Asia [Andamanese] can have introgression from same hominin populations having a similar amount of contribution [2%]). After revisiting, we realised searching for less African allele is more difficult than anticipated and the pipeline should be proper to detect less African ancestry. We demonstrated Andamanese showed less African allele compared to Europeans in every possible way we could think off. We reproduced the result in The 1000 Genomes Project data also (using Indian instead of Andamanese). It looks like having fewer individuals from the same population sometimes reduce the power to detect the absence of African ancestry, which might be the case for us not detecting this unknown hominin introgression in our East Asian samples. Nonetheless, this exercise improved our model of introgression and now simulated data perfectly matches with empirical data and we can solve a lot of controversy around archaic introgression to modern humans. We hypothesised that all Asian populations had introgression from this unknown hominin population once before they separated from each other. Therefore, in the end, OOA modern human populations had three introgressions in total from hominin populations: first Neanderthal introgressed in all OOA populations before they have separated, second this unknown population introgressed after

Asian populations have separated from Europeans and in the end, Pacific populations have introgression from Denisova after they have separated from Asian populations.

4.4 Hominin Introgression of Aboriginal Australians

Aboriginal Australians are one of the most important populations to understand how modern humans spread around the world. Right now, they are even more interesting as the controversy surrounding first OOA event for modern humans getting heated up. If our previous analysis were right, Aboriginal Australians should have similar ancestry like other Asian populations. This, in turn, makes them highly probable of having introgression from this unknown hominin population, which we found in other Asian populations. It is particularly complicated to find the signature of this unknown hominin inside Aboriginal Australians genome, as they also have introgression from another hominin population (Denisova). We used an improved way (by masking possible Denisova introgressed regions) to find this unknown hominin introgression using D-stats. We are able to show that Aboriginal Australians do have introgression from this unknown hominin population, indirectly suggesting again that they have similar ancestry like all other Asian populations.

4.5 Y-Chromosome Dilemma Solved for Andamanese

Lastly, we concentrated on Y chromosome analysis of Indian populations. We recapitulated some of the already known results: like in the North we have higher frequency of R haplogroup, which is also present in higher frequency in Europeans, most probably brought by the Aryan migration; Indian tribal population have few haplogroups which are not present outside of India; Tibeto-Burman populations having O haplogroup which is present in high frequency in East Asia and also Bengalis most probably because of proximity with Tibeto-Burman populations. The main interesting point of this paper was to look deep inside Andamanese D haplogroup. This haplogroup can only be detected in Andaman and Japan (and with some other part in the Himalaya all of them are remote places). This haplogroup has different origin compared to all other OOA haplogroups. This is not expected as we found that Andamanese has similar ancestry with Asian populations. Therefore, we did some simulations, showed that this Haplogroup D is most probably a standing variation when OOA event happened. This haplogroup was present in other OOA populations but disappeared because of drift. Only Andamanese and some Japanese individuals retained this haplogroup. In this instance, we proposed that Y-chromosome does not give the real ancestry of the population. Sometimes much older standing variation can survive by simple random chance in Y chromosome.

4.6 Epilogue

We finally reached the end of my thesis. Human population genetics studies are really exhilarating, as we start to understand our own history which gives a new perspective of our ancestor who failed to leave any written document - thus more likely to be unbiased. As Dan Brown put it correctly, “History is always written by the winners. When two cultures clash, the loser is obliterated, and the winner writes the history books-books which glorify their own cause and disparage the conquered foe.” Of course, population genetic studies cannot directly give insight on culture or spoken language but we can guess it from population genetic studies. In a nutshell, population genetic studies can give an insight only when there is sex involved and those individuals left some progenies. Although this might look primitive from orthodox people’s point of view, it is indeed a very powerful tool.

In this thesis, we concentrated on Indian populations as well as a little bit on Aboriginal Australian populations. These populations are generally underrepresented in population genetic studies till now, which gave us an edge to find this unknown hominin population. This piece of information has met with criticism as it was expected. Science is right now biased towards already known big players in the field, which put new budding scientist to a possible disadvantageous point (even for publishing in big journals). Nonetheless, a truth is a truth and will come out in the end. Unlike politics and arts where popular belief can change an outcome, in science that does not happen. We will eventually see who is right irrespective of the comments from the big players.

5. Abbreviations (except Methods and Results, Appendix)

ANI = Ancestral North Indian
ASI = Ancestral South Indian
BLAST = Basic Local Alignment Search Tool
BWA = Burrows-Wheeler Aligner
DNA = Deoxyribonucleic acid
D-stats = D-statistics
EM = Expectation–Maximization
GATK = Genome Analysis ToolKit
GDP = Gross Domestic Product
HB = Hierarchical Boosting
kya = kilo years ago
LD = Linkage Disequilibrium
MSMC = Multiple Sequentially Markovian Coalescent
OOA = Out of Africa
PC = Principal Component
PCA = Principal Component Analysis
PCR = Polymerase Chain Reaction
PhD = Doctor of Philosophy
PSMC = Pairwise Sequentially Markovian Coalescent
RAXML = Randomized Axelerated Maximum Likelihood
RBC = Red Blood Cell
SNP = Single Nucleotide Polymorphisms
TMRCA = The Most Recent Common Ancestor

6. Bibliography

Abbi,A. (2009) Is Great Andamanese genealogically and typologically distinct from Onge and Jarawa? *Lang. Sci.*, **31**, 791–812.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–10.

Ambrose,S.H. (1998) Late Pleistocene human population bottlenecks, volcanic winter, and differentiation of modern humans. *J. Hum. Evol.*, **34**, 623–651.

Armitage,S.J. *et al.* (2011) The Southern Route ‘Out of Africa’: Evidence for an Early Expansion of Modern Humans into Arabia. *Science* , **453**, 453–156.

Avari,B. (2007) India: The Ancient Past: A History of the Indian Sub-Continent from C. 7000 BC to AD 1200 Routledge.

Basu,A. *et al.* (2015) Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. *Pnas*, **113**, 201513197.

Bryant,E. (2003) The Quest for the Origins of Vedic Culture: The Indo-Aryan Migration Debate.

Casals,F. and Bertranpetit,J. (2012) Human Genetic Variation, Shared and Private. *Science* , **337**, 39–40.

Diamond,J. (2014) The Third Chimpanzee for Young People: On the Evolution and Future of the Human Animal.

Fay,J.C. and Wu,C.I. (2000) Hitchhiking under positive Darwinian selection. *Genetics*, **155**, 1405–1413.

Fisher,R.A. (1930) THE GENETICAL THEORY OF NATURAL SELECTION.

Fu,Y.X. and Li,W.H. (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.

Government of India (2011) Census of India 2011. *State Lit.*, 3–4.

Gravel,S. *et al.* (2011) Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 11983–11988.

Green,R.E. *et al.* (2010) A draft sequence of the Neandertal genome. *Science*, **328**, 710–722.

Grün,R. *et al.* (2005) U-series and ESR analyses of bones and teeth relating to the human burials from Skhul. *J. Hum. Evol.*, **49**, 316–334.

Gutenkunst,R.N. *et al.* (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.*, **5**, e1000695.

Hardy,G.H. (1908) Mendelian Proportions in a Mixed Population. *Science* , **28**, 49–50.

Huxley,T.H. (1870) On the Geographical Distribution of the Chief Modifications of Mankind. *J. Ethnol. Soc. London*, **2**, 404–412.

IUCN (2015) IUCN Red List of Threatened Species. *Version 2015.2*, www.iucnredlist.org.

Jorde,L.B. and Wooding,S.P. (2004) Genetic variation, classification and ‘race’. *Nat. Genet.*, **36**, S28-33.

Kimura,M. (1968) Evolutionary Rate at the Molecular Level. *Nature*, **217**, 624–626.

Kuhlwilm,M. *et al.* (2016) Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature*.

Lamason,R.L. *et al.* (2005) SLC24A5, a Putative Cation Exchanger, Affects Pigmentation in Zebrafish and Humans. *Science* , **310**, 1782–1786.

Lander,E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Langergraber,K. and Prüfer,K. (2012) Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc. ...*, **109**, 15716–15721.

Lawson,D.J. *et al.* (2011) Inference of Population Structure using Dense Haplotype Data. **8**, 11–17.

de León,M.S.P. *et al.* (2008) Neanderthal brain size at birth provides insights into the evolution of human life history. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 13764–13768.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. and Durbin,R. (2011) Inference of human population history from individual whole-genome sequences. *Nature*, **475**, 493–496.

Li,H. and Durbin,R. (2012) Inference of Human Population History From Whole Genome Sequence of A Single Individual. *Nature*, **475**, 493–496.

Liu,X. *et al.* (2013) Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *PLoS One*, **8**.

Lomolino,M. V. (2005) Body size evolution in insular vertebrates: Generality of the island rule. *J. Biogeogr.*, **32**, 1683–1699.

McCulloch,S.D. and Kunkel,T.A. (2008) The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Res.*, **18**, 148–61.

McDougall,I. *et al.* (2005) Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature*, **433**, 733–6.

McKenna,A. *et al.* (2010) The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

Melé,M. *et al.* (2012) Recombination gives a new insight in the effective population size and the history of the old world human populations. *Mol. Biol. Evol.*, **29**, 25–30.

Meyer,M. *et al.* (2012) A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* , **338**, 222–226.

Moorjani,P. *et al.* (2016) Variation in the molecular clock of primates. *bioRxiv Prepr.*, 1–39.

Nielsen,R. *et al.* (2005) Genomic scans for selective sweeps using SNP data. *Genome Res.*, **15**, 1566–1575.

Parton,A. *et al.* (2015) Alluvial fan records from southeast Arabia reveal multiple windows for human dispersal. *Geology*, **43**, 295–298.

Patterson,N. *et al.* (2012) Ancient admixture in human history. *Genetics*, **192**, 1065–1093.

Patterson,N., Richter,D.J., *et al.* (2006) Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, **441**, 1103–8.

Patterson,N., Price,A.L., *et al.* (2006) Population structure and eigenanalysis. *PLoS Genet.*, **2**, 2074–2093.

Pickrell,J.K. and Pritchard,J.K. (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.*, **8**, e1002967.

Poznik,G.D. *et al.* (2016) Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat. Genet.*, **12**, 809–809.

Prado-Martinez,J. *et al.* (2013) Great ape genetic diversity and population history. *Nature*,

499, 471–475.

Pybus,M. *et al.* (2015) Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. *Bioinformatics,* **31**, 3946–52.

Raia,P. and Meiri,S. (2006) The island rule in large mammals: paleontology meets ecology. *Evolution*, **60**, 1731–1742.

Rasmussen,M. *et al.* (2011) An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. *Science* , **334**, 94–98.

Reich,D. *et al.* (2009) Reconstructing Indian population history. *Nature*, **461**, 489–494.

Sabeti,P.C. *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature,* **419**, 832–837.

Sabeti,P.C. *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–918.

Sabeti,P.C. *et al.* (2006) Positive natural selection in the human lineage. *Science* , **312**, 1614–1620.

Scerri,E.M.L. *et al.* (2014) Earliest evidence for the structure of Homo sapiens populations in Africa. *Quat. Sci. Rev.*, **101**, 207–216.

Schadt,E.E. *et al.* (2010) A window into third-generation sequencing. *Hum. Mol. Genet.,* **19**.

Schiffels,S. and Durbin,R. (2014) Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.*, **46**, 919–925.

Smith,T.M. *et al.* (2007) Earliest evidence of modern human life history in North African early Homo sapiens. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 6128–6133.

Soares,P. *et al.* (2009) Correcting for Purifying Selection: An Improved Human Mitochondrial Molecular Clock. *Am. J. Hum. Genet.*, **84**, 740–759.

Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics,* **30**, 1312–3.

Suwa,G. *et al.* (2007) A new species of great ape from the late Miocene epoch in Ethiopia. *Nature*, **448**, 921–4.

Tajima,F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics,* **123**, 585–595.

Van Valen,L. (1973) A new evolutionary theory. *Evol. Theory*, **1**, 1–30.

Venter,J.C. *et al.* (2001) The sequence of the human genome. *Science,* **291**, 1304–51.

Vernot,B. *et al.* (2014) Resurrecting surviving Neandertal lineages from modern human genomes. *Science* , **343**, 1017–1021.

Voight,B.F. *et al.* (2006) A map of recent positive selection in the human genome. *PLoS Biol.*, **4**, 0446–0458.

Weir,B.S. and Hill,W.G. (2002) Estimating F-Statistics. *Annu. Rev. Genet,* **36**, 721–50.

White,T.D. *et al.* (2009) Ardipithecus ramidus and the paleobiology of early hominids. *Science*, **326**, 75–86.

Wolpoff,M.H. *et al.* (2000) Multiregional, not multiple origins. *Am. J. Phys. Anthropol.*, **112**, 129–136.

Wright,S. (1931) Evolution in Mendelian Populations. *Genetics*, **16**, 97–159.

Zahavi,A. (1975) Mate selection-A selection for a handicap. *J. Theor. Biol.*, **53**, 205–214.

7. Appendix

7.1 Supplementary Information of "Population and genomic lessons from genetic analysis of two Indian populations".

Juyal G, Mondal M, Luisi P, Laayouni H, Sood A, Midha V, et al. Population and genomic lessons from genetic analysis of two Indian populations. Supplementary material. Hum Genet. 2014 Oct 1;133(10):1273–87. DOI: 10.1007/s00439-014-1462-0

7.2 Supplementary Information of "Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation."

With slight changes for the positions of the graphs to increase readability from the real version which is available online.

Mondal M, Casals F, Xu T, Dall'Olio GM, Pybus M, Netea MG, et al. Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. Supplementary material. Nat Genet. 2016 Sep 25;48(9):1066–70. DOI: 10.1038/ng.3621