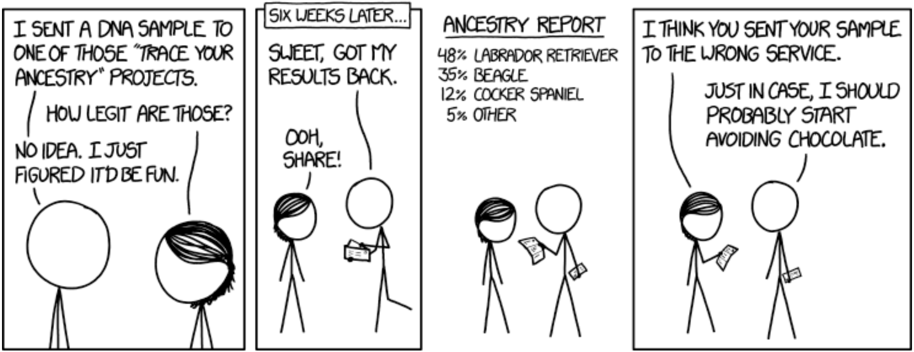

I SENT A DNA SAMPLE TO ONE OF THOSE "TRACE YOUR ANCESTRY" PROJECTS.
HOW LEGIT ARE THOSE?
NO IDEA. I JUST FIGURED IT'D BE FUN.
SIX WEEKS LATER...
SWEET, GOT MY RESULTS BACK.
OOH, SHARE!
ANCESTRY REPORT
48% LABRADOR RETRIEVER
35% BEAGLE
12% COCKER SPANIEL
5% OTHER
I THINK YOU SENT YOUR SAMPLE TO THE WRONG SERVICE.
JUST IN CASE, I SHOULD PROBABLY START AVOIDING CHOCOLATE.

https://xkcd.com/1706/

Simons Genome Diversity Project

Don L. Armstrong

Institute for Genomic Biology, Computing Genomes for Reproductive Health, University of Illinois, Urbana-Champaign

October 20, 2016

Code and slides are here:

`http://dla2.us/p/genomdiv2016`

doi:10.1038/nature18964

The Simons Genome Diversity Project: 300 genomes from 142 diverse populations

A list of authors and affiliations appears at the end of the paper.

Here we report the Simons Genome Diversity Project data set: high quality genomes from 300 individuals from 142 diverse populations. These genomes include at least 5.8 million base pairs that are not present in the human reference genome. Our analysis reveals key features of the landscape of human genome variation, including that the rate of accumulation of mutations has accelerated by about 5% in non-Africans compared to Africans since divergence. We show that the ancestors of some pairs of present-day human populations were substantially separated by 100,000 years ago, well before the archaeologically attested onset of behavioural modernity. We also demonstrate that indigenous Australians, New Guineans and Andamanese do not derive substantial ancestry from an early dispersal of modern humans; instead, their modern human ancestry is consistent with coming from the same source as that of other non-Africans.

To obtain a complete picture of human diversity, it is necessary to sequence the genomes of many individuals from diverse locations. To date, the largest whole-genome sequencing survey, the 1000 Genomes Project, analysed 26 populations of European, East Asian, South Asian, American, and sub-Saharan African ancestry[1]. However, this and most other sequencing studies have focused on demographically large populations. Such studies tend to ignore smaller populations that are

required for polymorphism discovery and analysis, and identified SNPs by comparing against the human reference. We find that FermiKit has comparable sensitivity and specificity to GATK for SNP discovery and genotyping, and is more accurate for indels (Supplementary Information section 4). FermiKit also identified 5.8 Mb of contigs that are present in the SGDP but absent in the human reference genome presumably because they are deleted there; these contigs which we

Sampling

- 142 populations from Africa, America, Oceania, South Asia, East Asia, and West Eurasia (mostly indigenous)
- 300 samples sequenced at 34-83 fold coverage by Illumina
- Aligned using BWA-MEM
- Genotyped using special version of GATK and Fermikit
- Data available in EBI ($n = 279$, PRJEB9586) and dbGAP ($n = 21$, ?)

Alignment Pipeline

- Aligned to the “decoy” version of the human reference (hs37d5); supposedly improves alignment in misassembled regions or regions with CNVs?
- PCR-free data, though they marked optical duplicates marked using samblaster

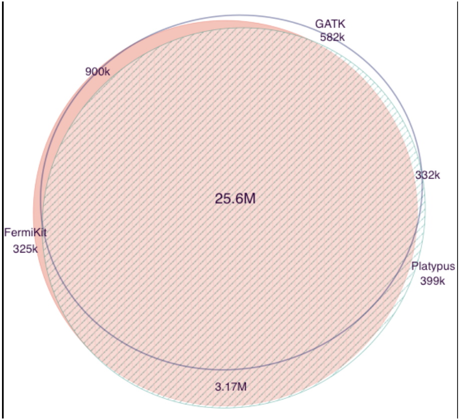
```
./htscmd bamshuf -Oun128 in.bam tmp-pre \
| ./htscmd bam2fq -as aln-se.fq.gz - \
| ./trimadap \
| ./bwa mem -pt8 hs37d5.fa - \
| ./samblaster \
| samtools view -uS - \
| samtools sort -@4 -m512M - out-pre
```

Genotyping

- Reference-bias; novel variants, GATK assumes reference is more likely which may not be the case. Use prior of $(0.4995, 0.001, 0.4995)$ instead of default $(0.9985, 0.001, 0.0005)$.
    - Unclear what the effect of this change is on the calling
    - Maybe worth thinking about?
- Also used Fermikit; apparently has comparable call rates to GATK and platypus

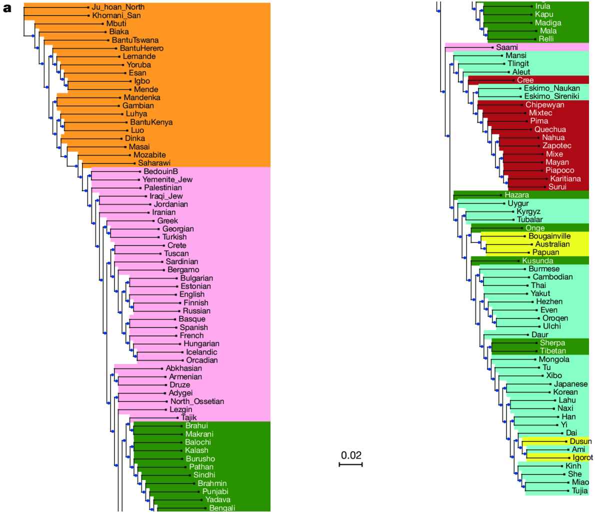
```
java -Xmx2g -jar GenomeAnalysisTK.jar \
-T UnifiedGenotyper -I srt.aln.bam \
-L CHR_ID -R hs37d5.fa -dcov 600 -glm SNP \
-out_mode EMIT_ALL_SITES -stand_call_conf 5.0 \
-stand_emit_conf 5.0 -inputPrior 0.0010 \
-inputPrior 0.4995 -D dbsnp_138.b37.vcf \
-o CHR_ID.vcf -A GCContent -A BaseCounts
```

Fermikit vs Platypus vs GATK



FermiKit and Platypus call 3.17M more sites than GATK, but unclear whether those are real sites or not; they go into this in much more detail than I've digested yet.

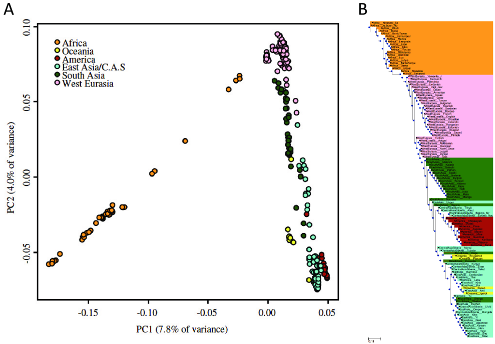
Relatedness of Populations



- Neighbor joining tree based on pairwise divergence per nucleotide
- Deepest splits are in African populations

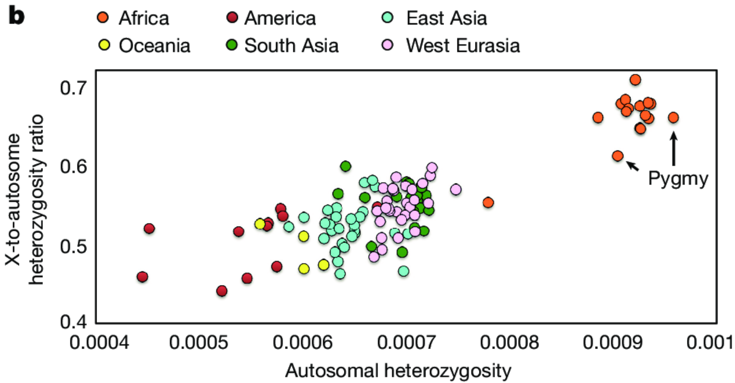
PCA and Relatedness

RESEARCH ARTICLE



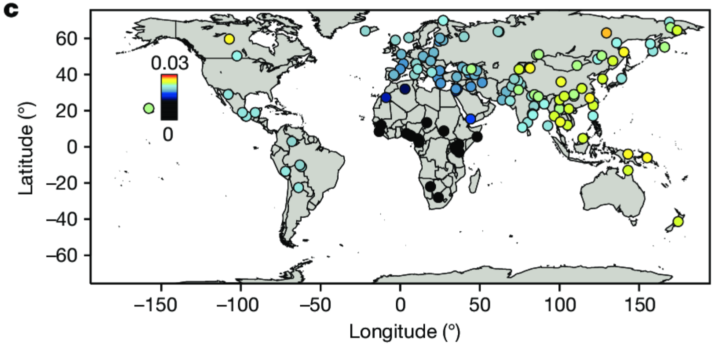
Extended Data Figure 4 | Principal component analysis and neighbour joining tree. a, Principal component analysis. **b**, Neighbour-joining tree based on F_{ST} values for all populations with at least two samples.

- Greatest variation seen in the African populations (orange)
- Other populations are much more similar to eachother in general
- Hapmap likely under-measured variation in Africa



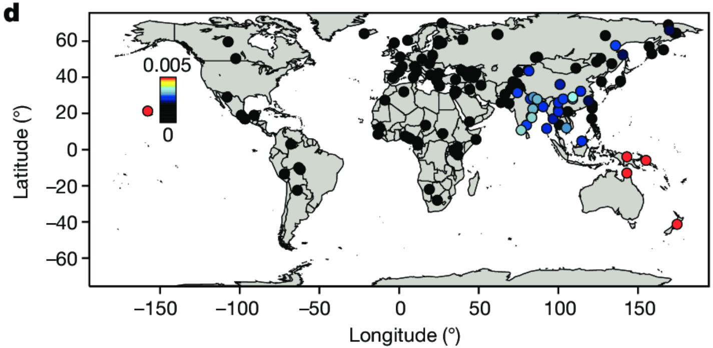
- Pygmy populations have lower X heterozygosity than other African populations
- Seen even after removing the third of X which is subject to selection
- Suggests that it’s driven by demographic history, and the reduced diversity is due to male-driven admixture (also in non-Africans)

Neanderthal Ancestry



- No populations studied have a higher Neanderthal ancestry than East Asians

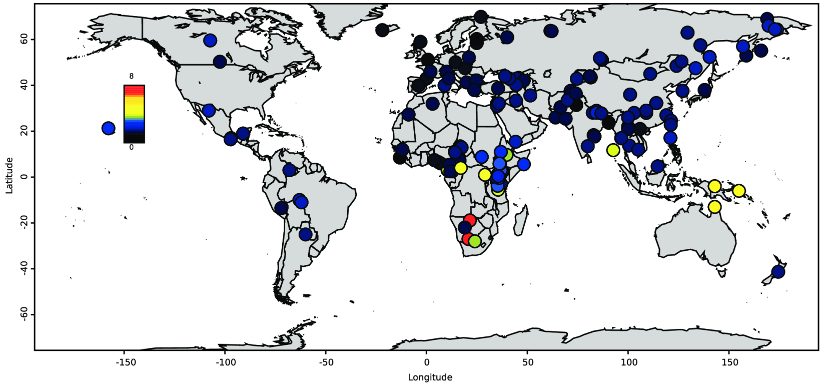
Denisovan Ancestry



- Oceanian groups have as much as 5% Denisovan ancestry
- Eurasian differences in ancestry; some South Asians may have higher Denisovan than other Eurasians

Variation missed by hapmap

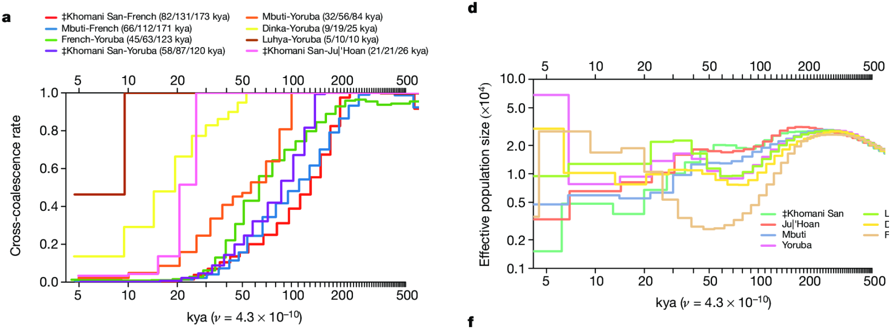
ARTICLE RESEARCH



Extended Data Figure 1 | Heat map of fraction of heterozygous sites missed in the 1000 Genomes project. For each sample, we examine all heterozygous sites passing filter level 1, and compute the fraction included as known polymorphisms in the 1000 Genomes project.

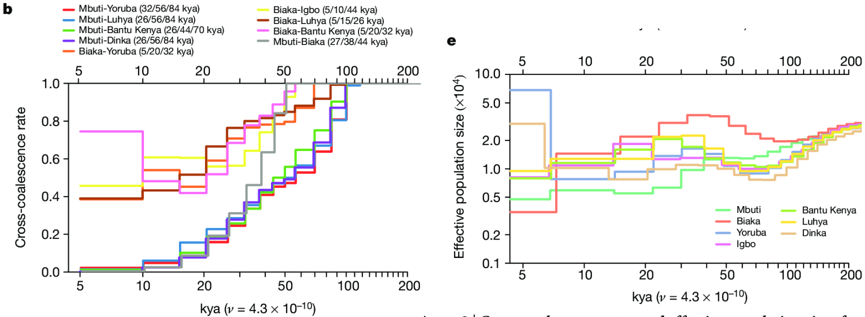
- Hapmap is missing up to 8% of the heterozygous sites in parts of Africa

Selected African cross-coalescence rate/Population Size



- Separation began around 200 kya for present-day hunter-gatherers
- Shared ancestors as recently as 100 kya

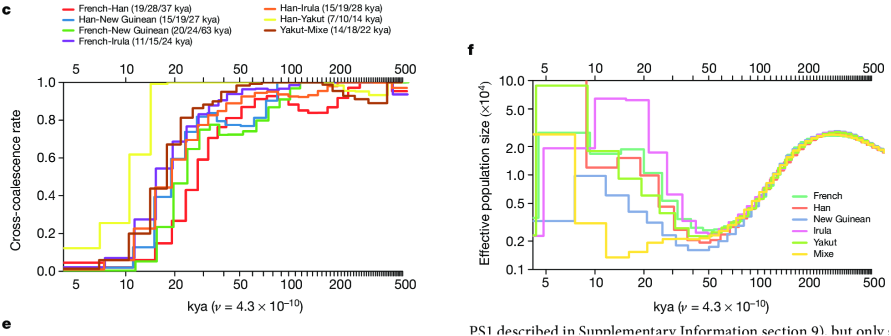
Central African rainforest hunter-gatherer



igure 2 | Cross-coalescence rates and effective population sizes for

- Within africa separation begins around 100 kya
- Still intermixing between Biaka (CAR, DRC), Bantu, and Luhya (Western Kenya)

Ancient, non-african



PS1 described in Supplementary Information section 9), but only

- Separation begins around 50 kya

Best-fitting admixture Graph

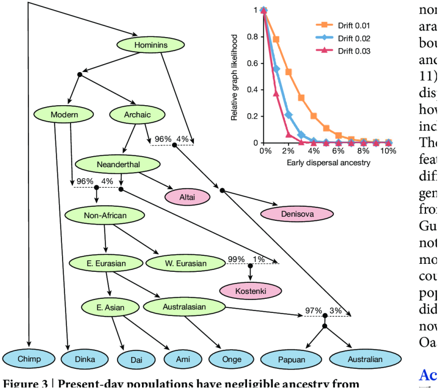


Figure 3 | Present-day populations have negligible ancestry from

noı
ara
bou
anc
11)
dis
hov
inc
The
feat
dif
gen
fro
Gu
not
mo
cou
pop
did
nov
Oa

Ac

- Present day populations have negligible ancestry from an early dispersal of modern humans out of africa

Conclusions

- Dispersal of human populations – little evidence of ancestry from an early dispersal of modern humans
- Possible acceleration in the rate of mutations among non-Africans
- No evidence for species-wide selective sweeps around ≈ 50 kya (start of modern human behavior in archaeological record) [Did not see that all pairs of modern humans shared a common ancestor 100 kya.
- Great resource for additional variants; maybe with H3A?
- Useful for my work with ancestral trees

References