Genomic epidemiology in Africa

Manjinder Sandhu

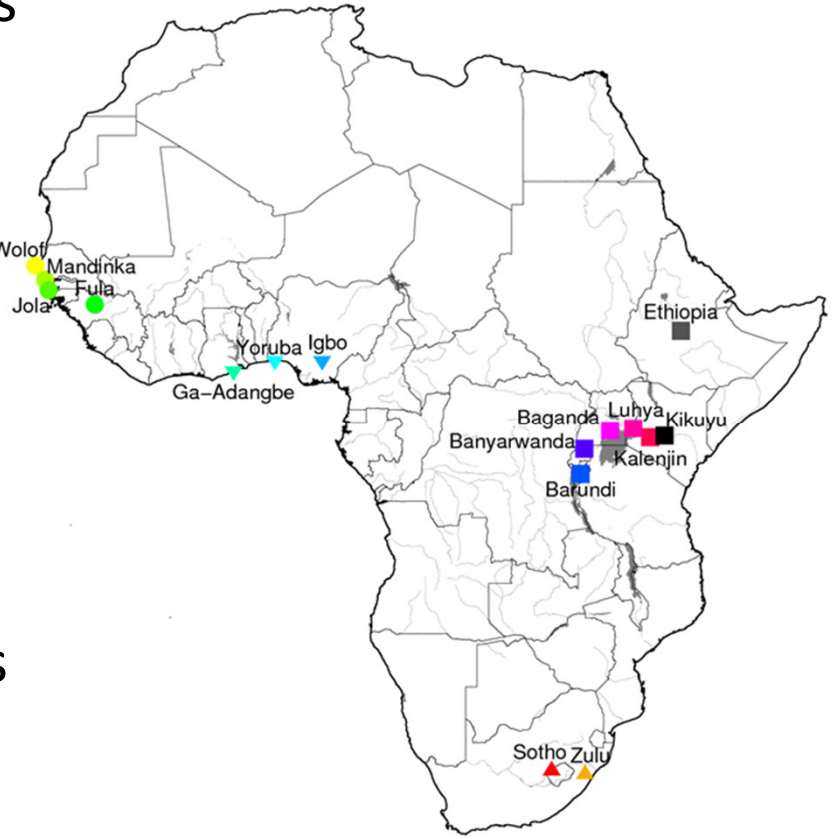
ms23@sanger.ac.uk

Genetic Epidemiology Group
Wellcome Trust Sanger Institute

International Health Research Group
Department of Public Health and Primary Care

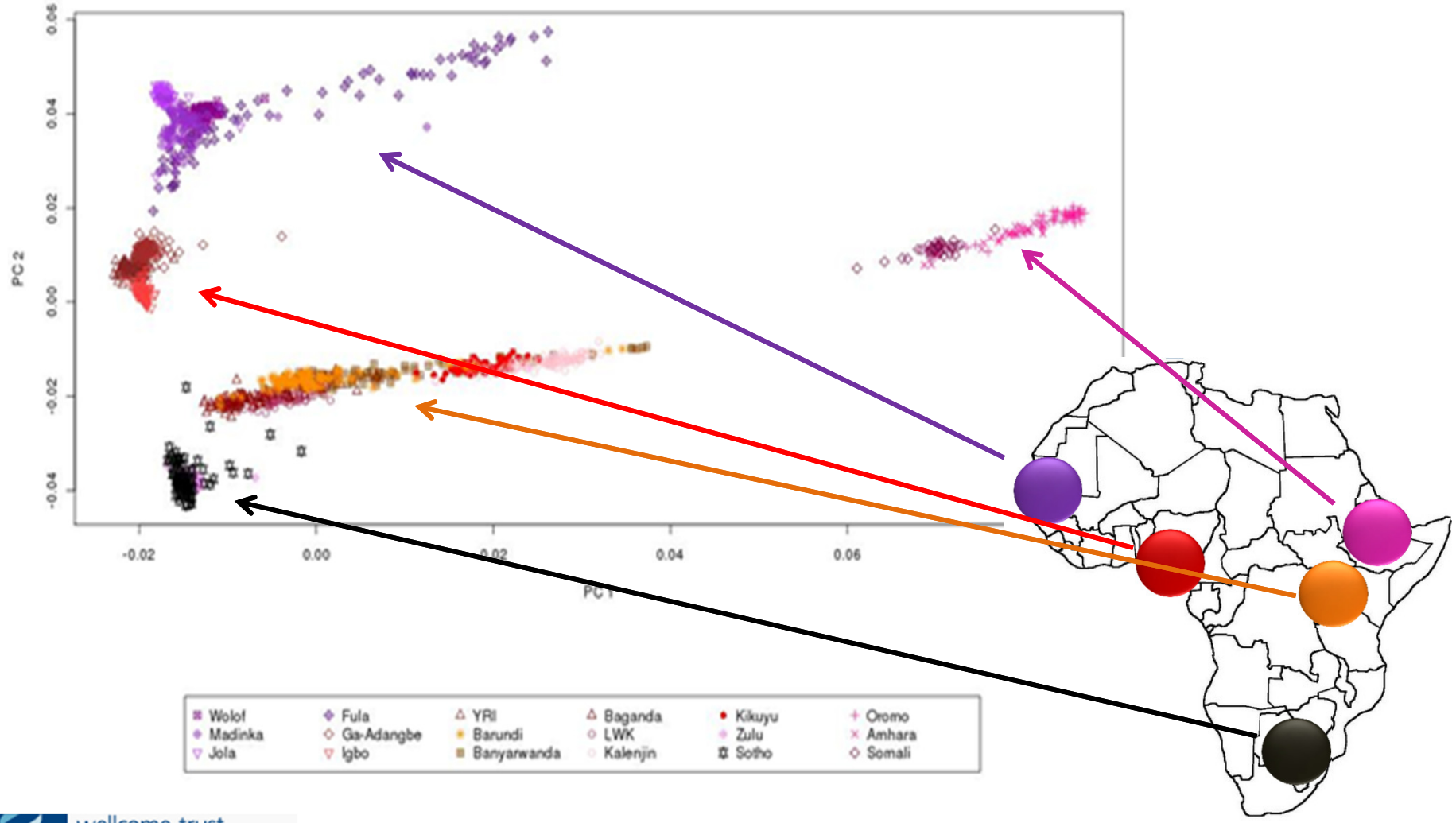
What is the African Genome Variation Project?

- Study of 16 ethno-linguistic groups across SSA from populations relevant to medical genomics
- 100 individuals with dense (2.5M) genotype data in each
- One of the largest diversity panel from Africa so far
- Aim: to study genetic variation in Africa to inform large scale studies in African genomics



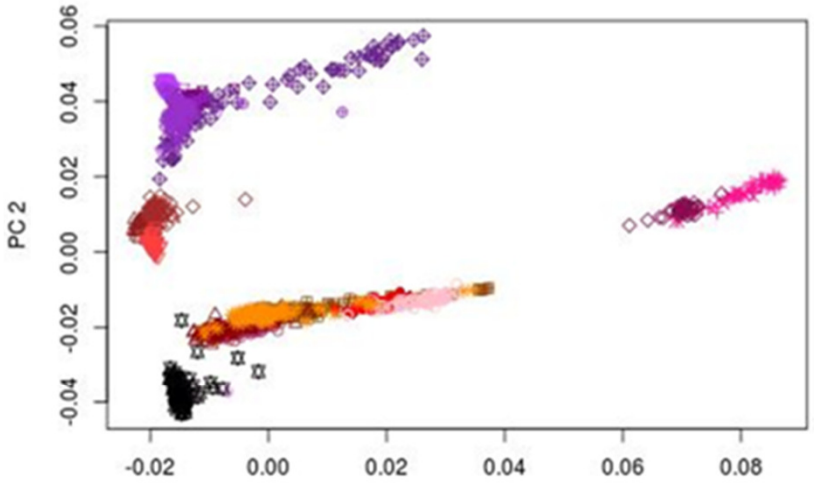
wellcome trust sanger institute

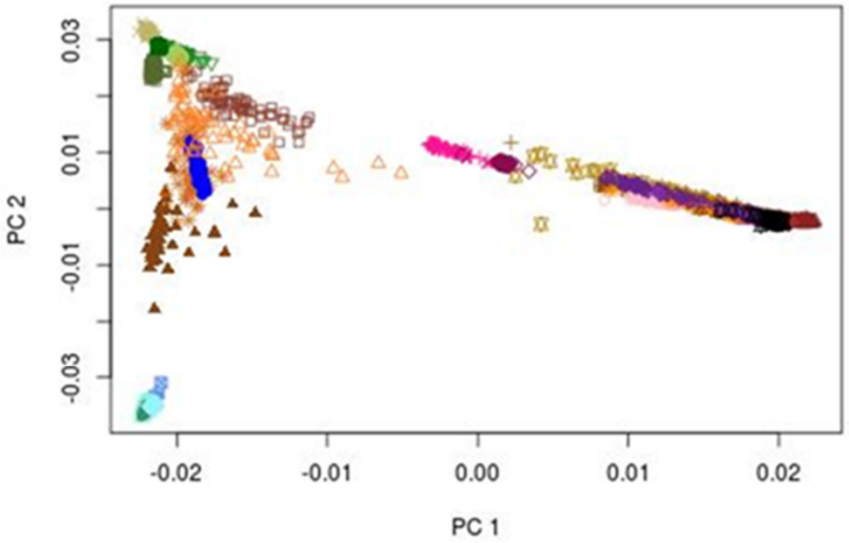
Population structure in SSA 1

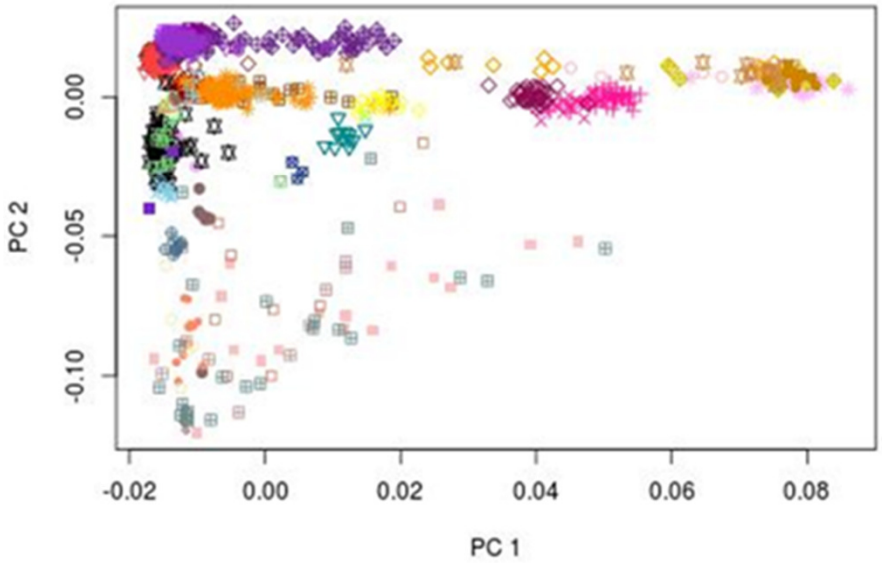


wellcome trust sanger institute

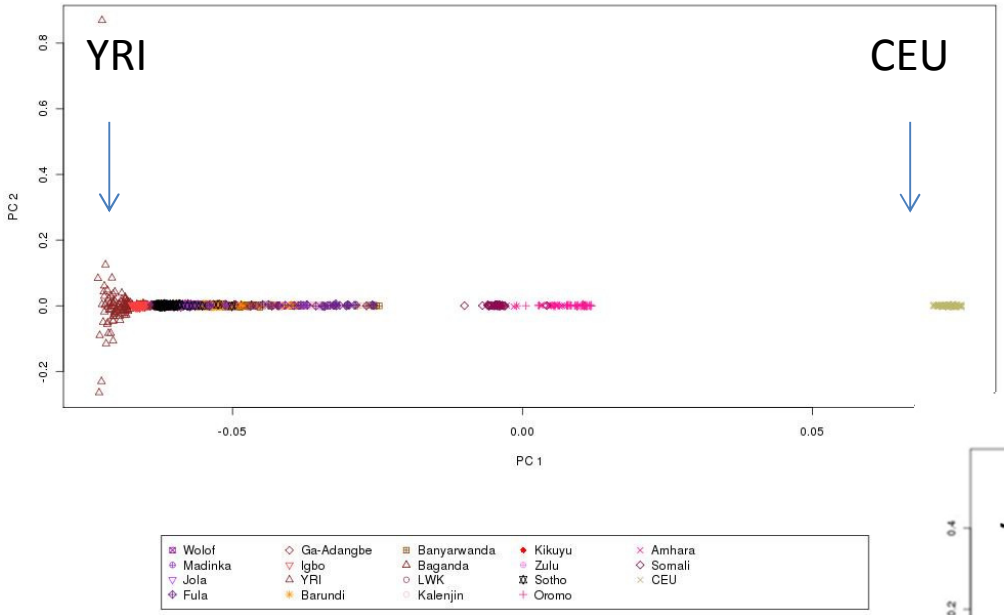
Population structure in SSA 3

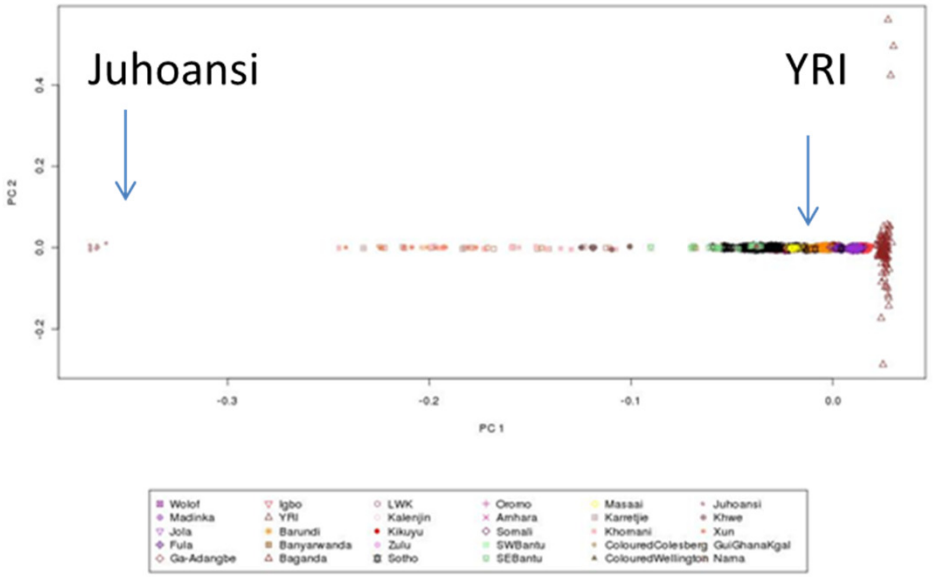






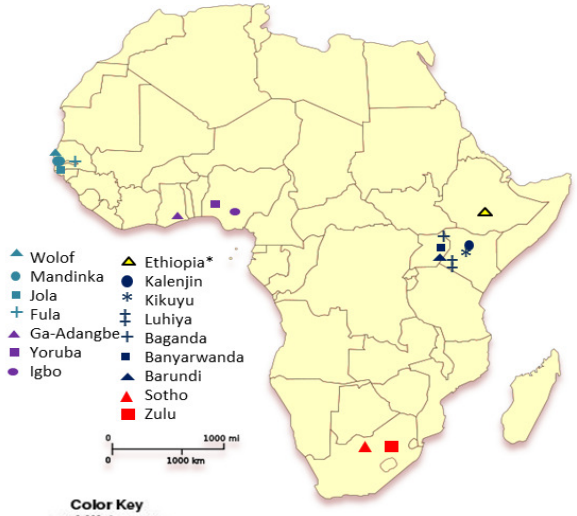
Evidence for Eurasian and KhoeSan admixture

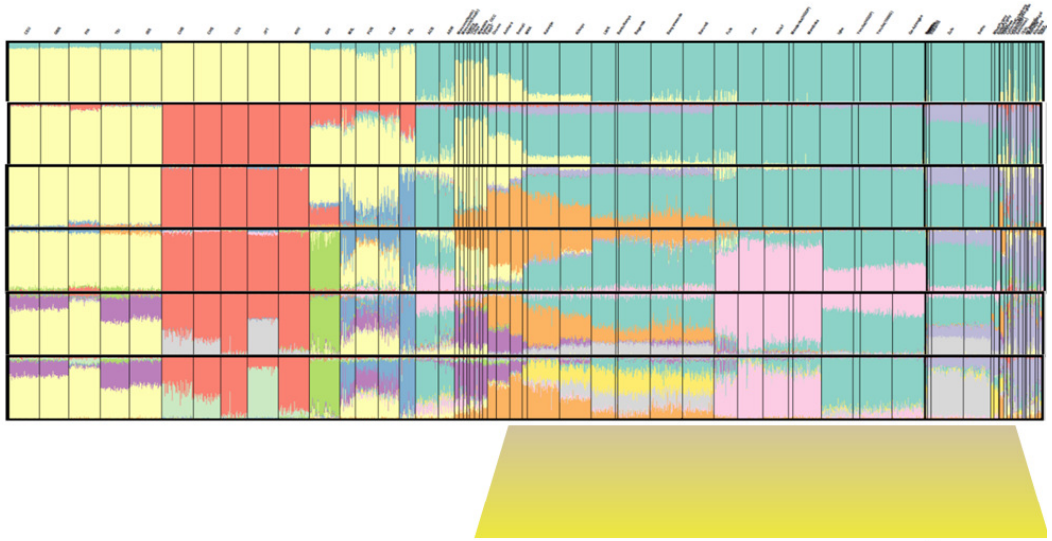


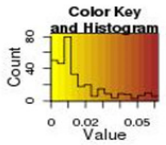


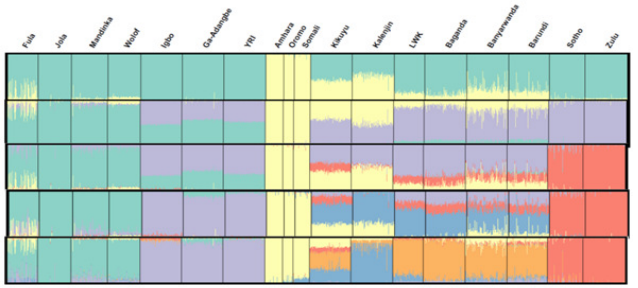
wellcome trust sanger institute

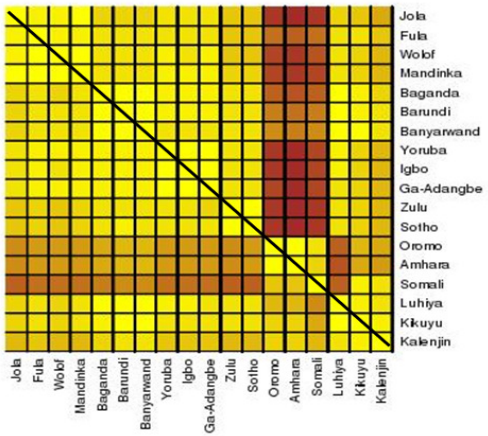
Population structure in SSA 2







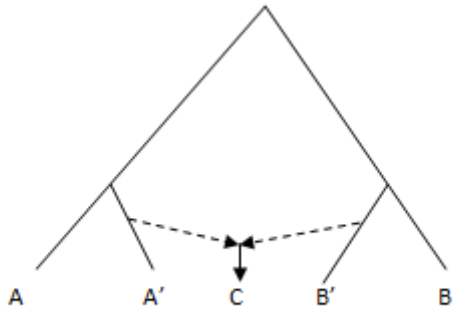


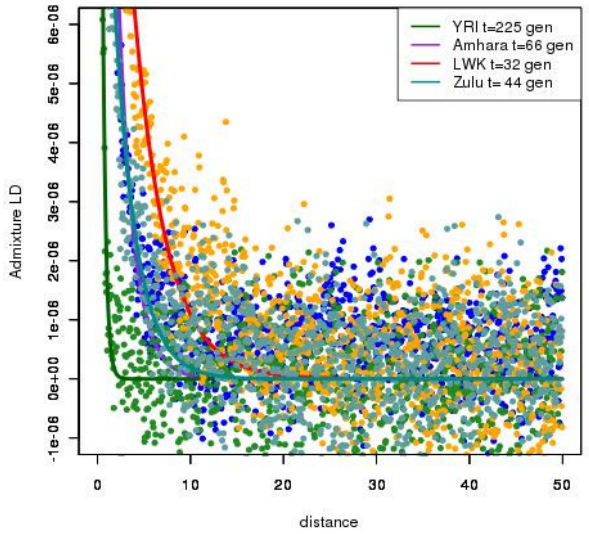


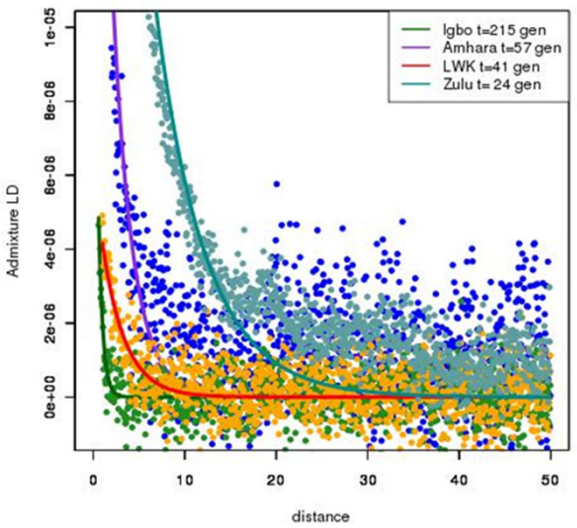
Confirming Eurasian and KhoeSan admixture

pop A	pop B	pop C	f3,C;A,B	Z score
CEU	YRI	Zulu	0.00543	33.528
CEU	YRI	Fula	-0.01115	-59.361
CEU	YRI	Jola	0.005366	31.897
CEU	YRI	Mandinka	-4.5E-05	-0.338
CEU	YRI	Wolof	-0.00079	-5.376
CEU	YRI	Baganda	-0.00316	-23.709
CEU	YRI	Banyarwanda	-0.01206	-74.707
CEU	YRI	Barundi	-0.00904	-64.071
CEU	YRI	LWK	-0.00322	-21.533
CEU	YRI	Kalenjin	-0.01042	-50.547
CEU	YRI	Kikuyu	-0.01618	-89.291
CEU	YRI	SOMALI	-0.01459	-47.375
CEU	YRI	AMHARA	-0.02385	-87.02
CEU	YRI	OROMO	-0.02391	-85.192
CEU	YRI	Igbo	0.000502	5.297
CEU	YRI	Ga-Adangbe	0.000008	0.078
CEU	YRI	Sotho	0.005022	31.282

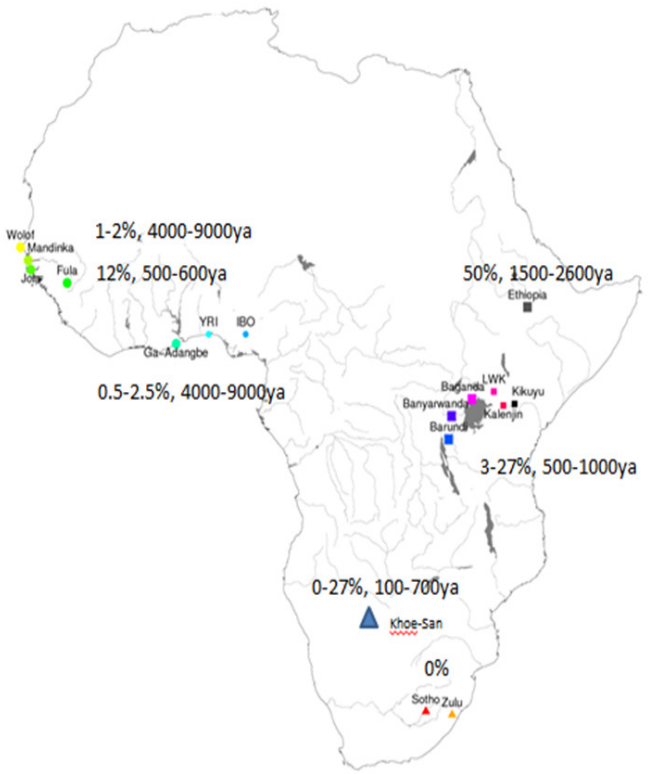
pop A	pop B	pop C	f3,C;A,B	Z score
YRI	Juhoansi	Zulu	-0.00705	-44.604
YRI	Juhoansi	Sotho	-0.00984	-64.773
YRI	Juhoansi	Fula	0.005064	38.379
YRI	Juhoansi	Jola	0.007087	51.022
YRI	Juhoansi	Mandinka	0.003206	29.069
YRI	Juhoansi	Wolof	0.005002	40.037
YRI	Juhoansi	Baganda	-0.00067	-6.157
YRI	Juhoansi	Barundi	-0.00207	-19.407
YRI	Juhoansi	Banyarwanda	-0.00044	-3.81
YRI	Juhoansi	LWK	-0.00017	-1.394
YRI	Juhoansi	Kikuyu	0.003081	20.333
YRI	Juhoansi	Kalenjin	0.009667	49.396
YRI	Juhoansi	Igbo	-0.0004	-5.39
YRI	Juhoansi	Ga-Adangbe	-5.9E-05	-0.773
YRI	Juhoansi	OROMO	0.042493	92.355
YRI	Juhoansi	AMHARA	0.048456	99.369
YRI	Juhoansi	SOMALI	0.043256	94.98

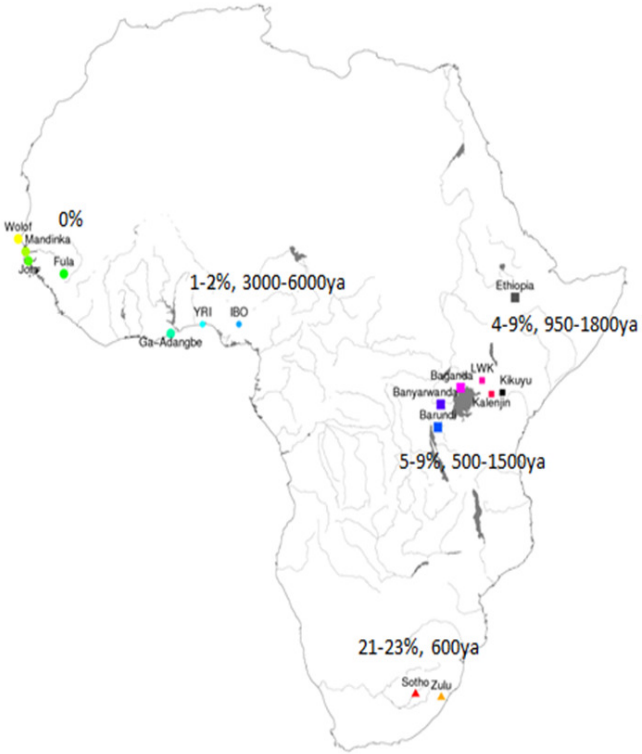




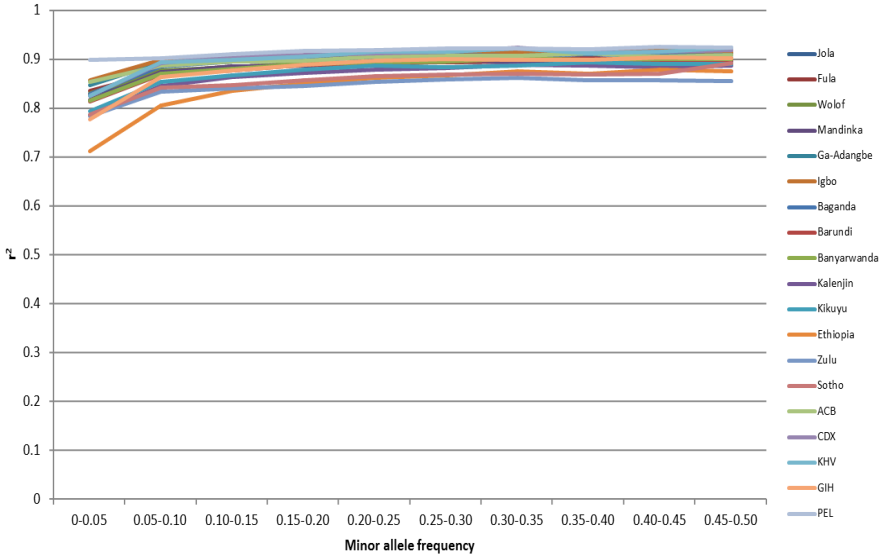


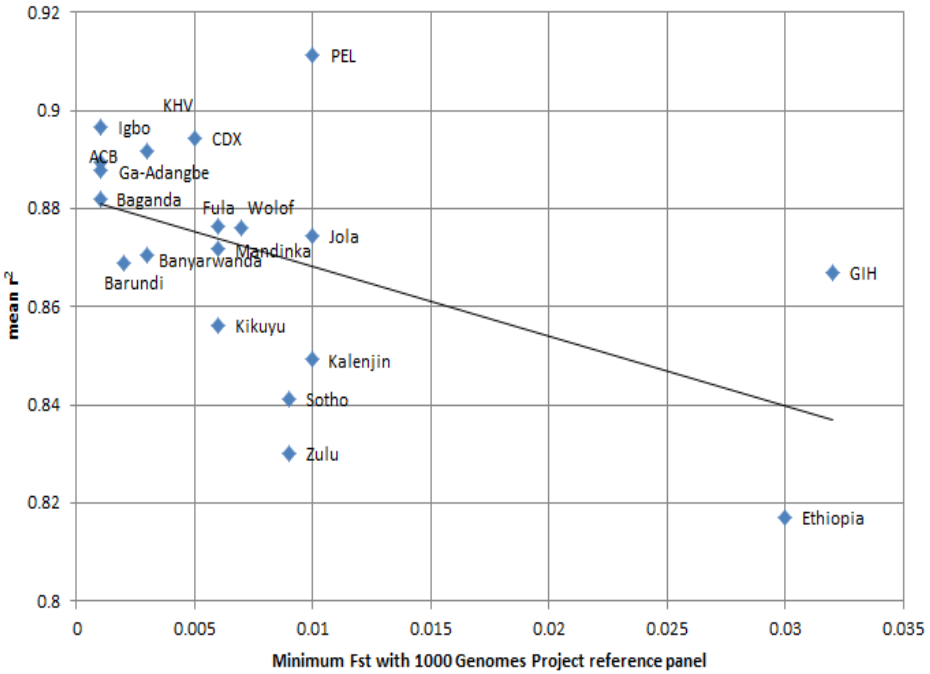
Quantifying and dating admixture in SSA



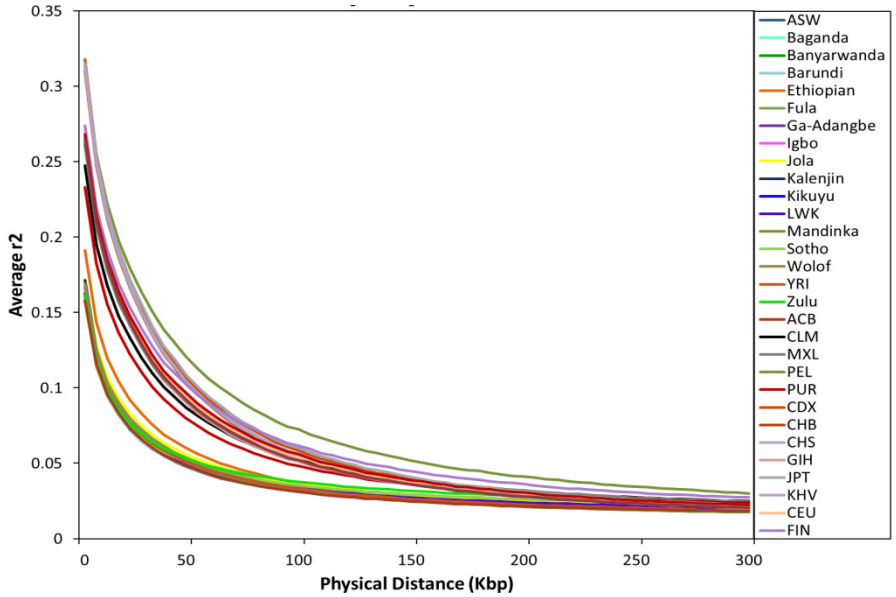


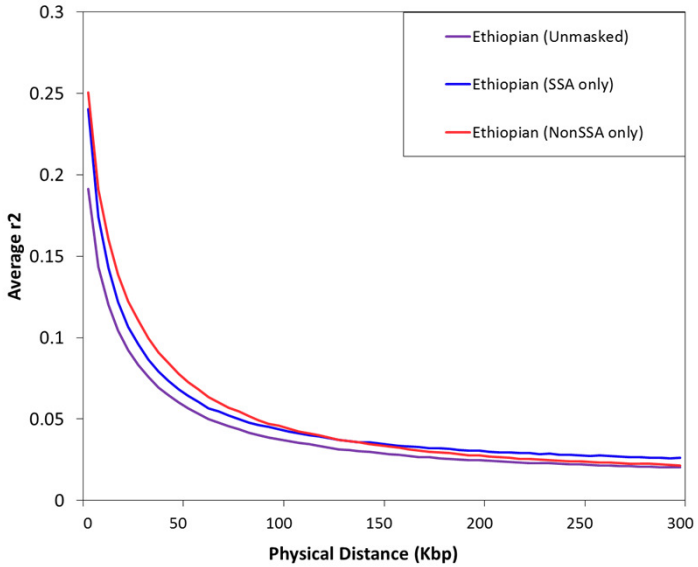
Imputation in SSA



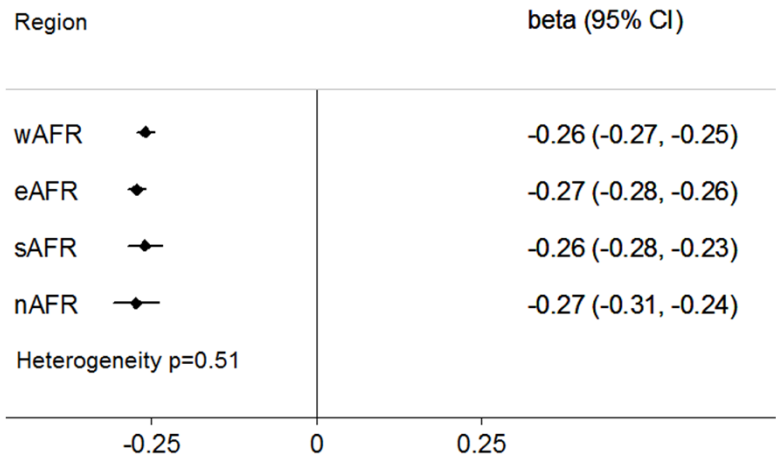


Linkage disequilibrium in SSA





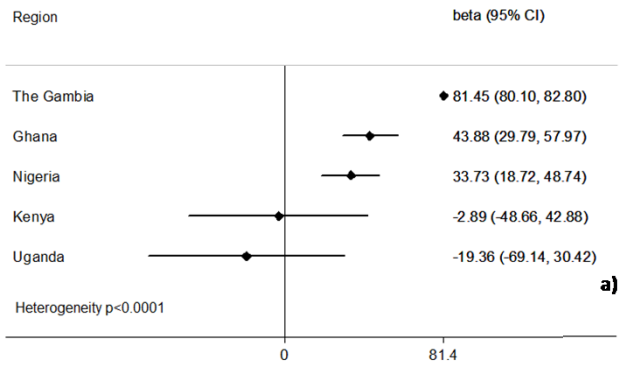
a) SORT1 locus (top candidate: rs646776 in all populations)



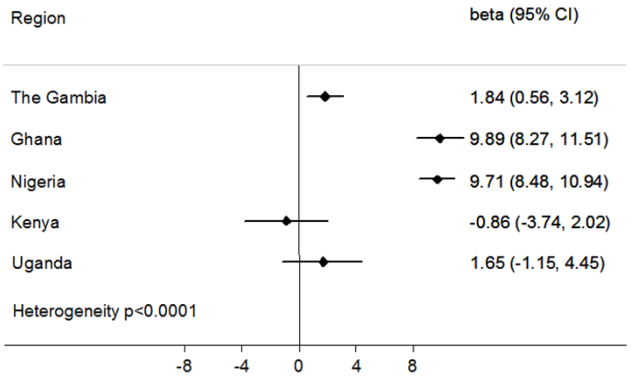
•Sickle cell locus (top candidate: rs113850170 in Kenya)

Linkage disequilibrium in SSA

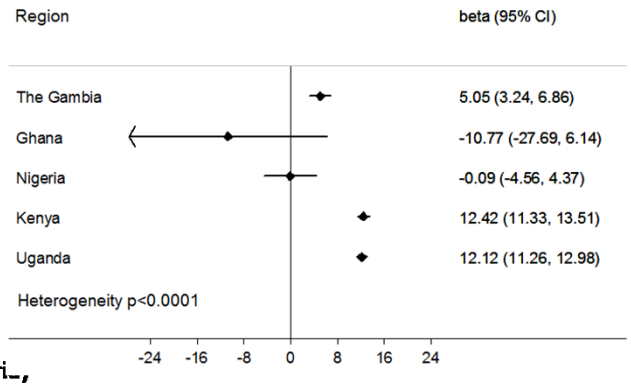
a) Sickle cell locus (top candidate: rs189784625 in the Gambia)



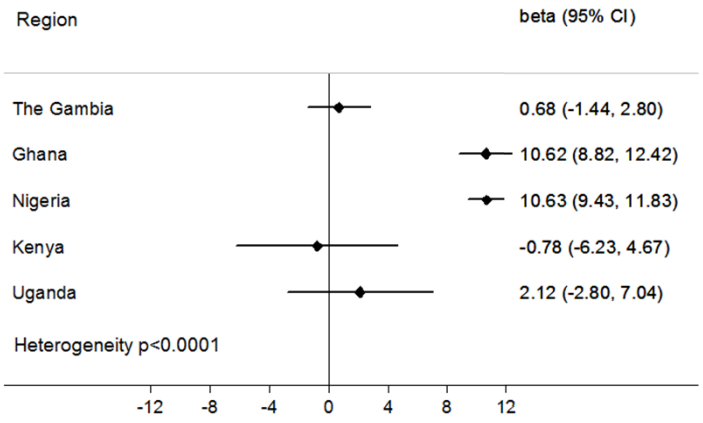
a) Sickle cell locus (top candidate: rs61735483 in Ghana)



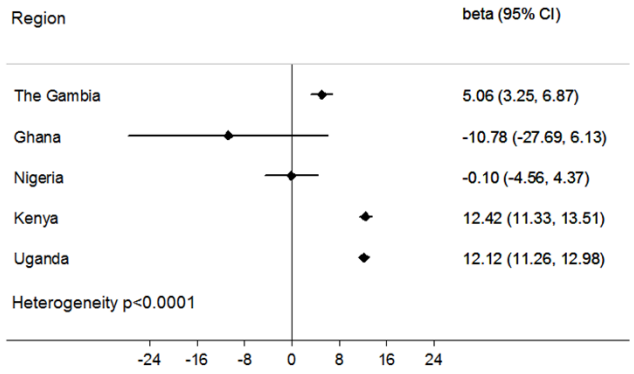
a) Sickle cell locus (top candidate: rs113850170 in Kenya)



a) Sickle cell locus (top candidate: chr11:5207604 in Nigeri_,



a) Sickle cell locus (top candidate: rs141845179 in Uganda)



Genome Diversity in Africa Project

- APCDR – 14 centres in 10 countries across Africa
- H3A Africa
- Wellcome Trust Sanger Institute
- MRC, Uganda

H3Africa
Human Heredity and Health in Africa

APCDR

MRC Medical Research Council

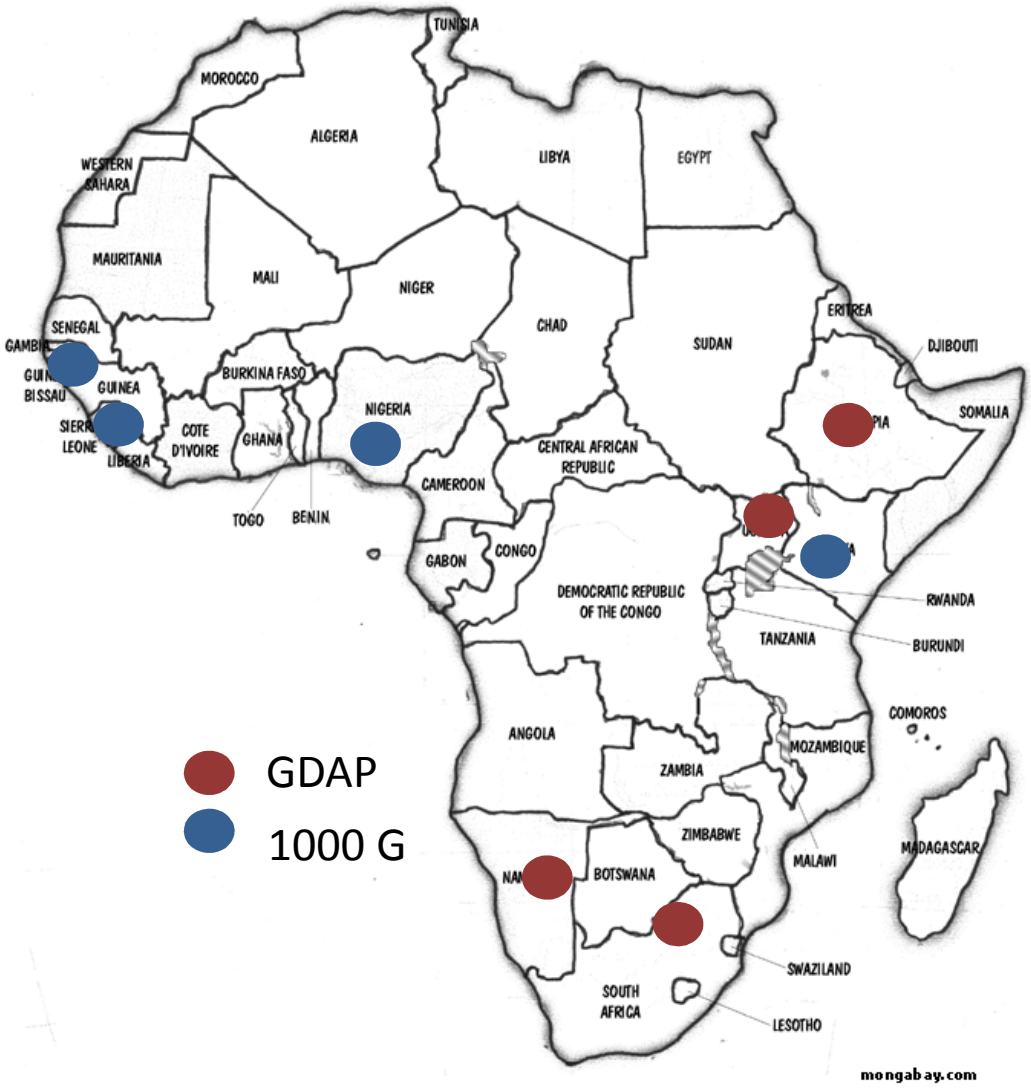
wellcome trust sanger institute

Key questions

- How do we best capture common and rare from African populations?
- What designs are best for large-scale genomic studies in Africa- how can these be optimised in terms of cost efficiency?
- How much of variation in African populations is private?
- Can data from different parts of Africa be combined?
- Can Africa specific reference panels improve imputation into African populations?
- How should such reference panels be curated in light of the variation between populations?
- Do we need an Africa specific chip to capture the variation across Africa?

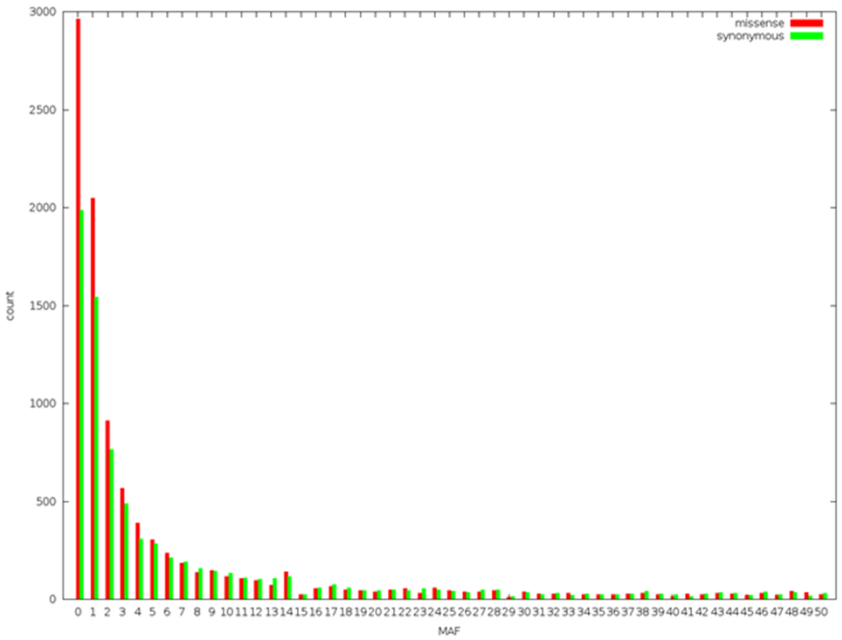
Sampling to date

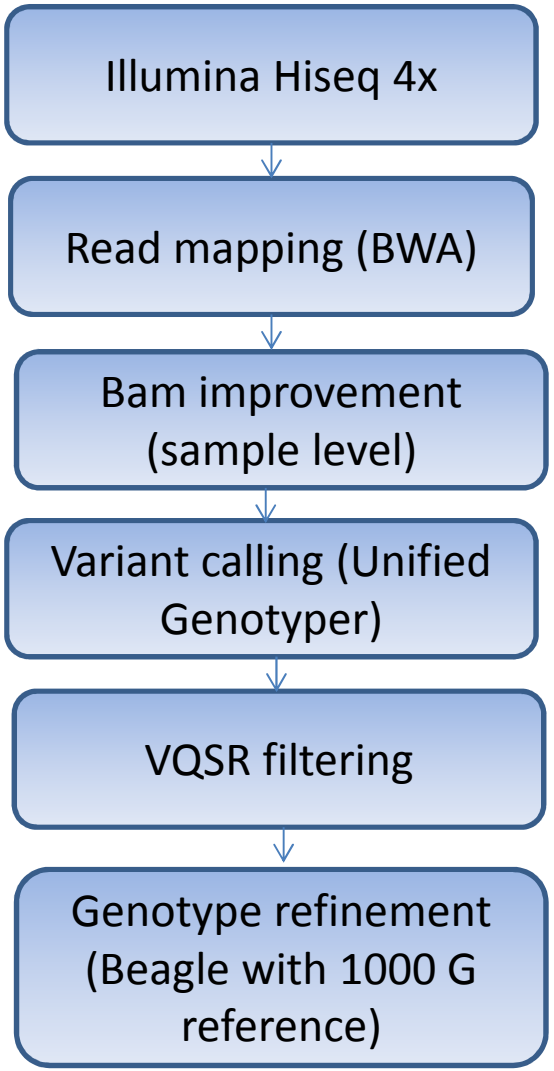
- 4x sequencing of 3 populations completed (320 samples)
- 2100 more in the pipeline
- Collections ongoing in East and South Africa



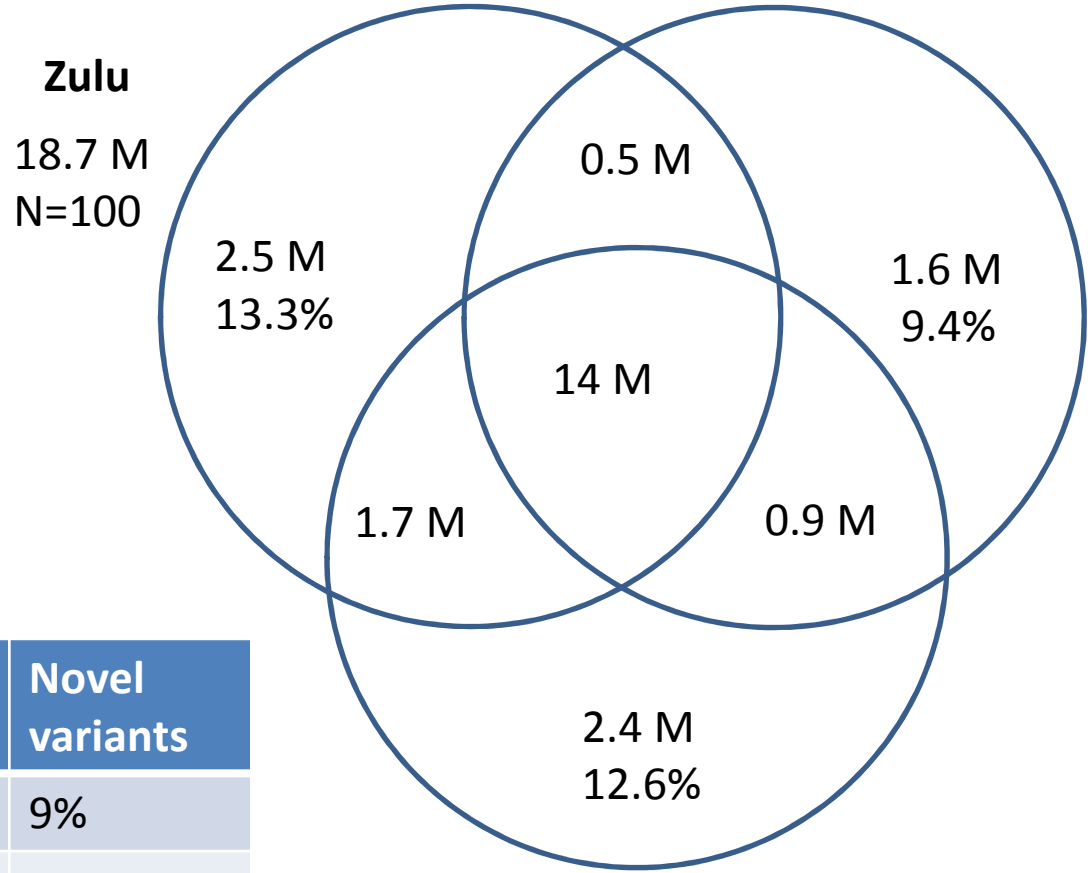
Data processing and output

Populations	r2	Concordance
Baganda	0.96	0.99
Ethiopia	0.95	0.99
Zulu	0.95	0.99





Preliminary results: sharing of variants



Pop	Novel variants
Zulu	9%
Baganda	7%
Ethiopia	15%

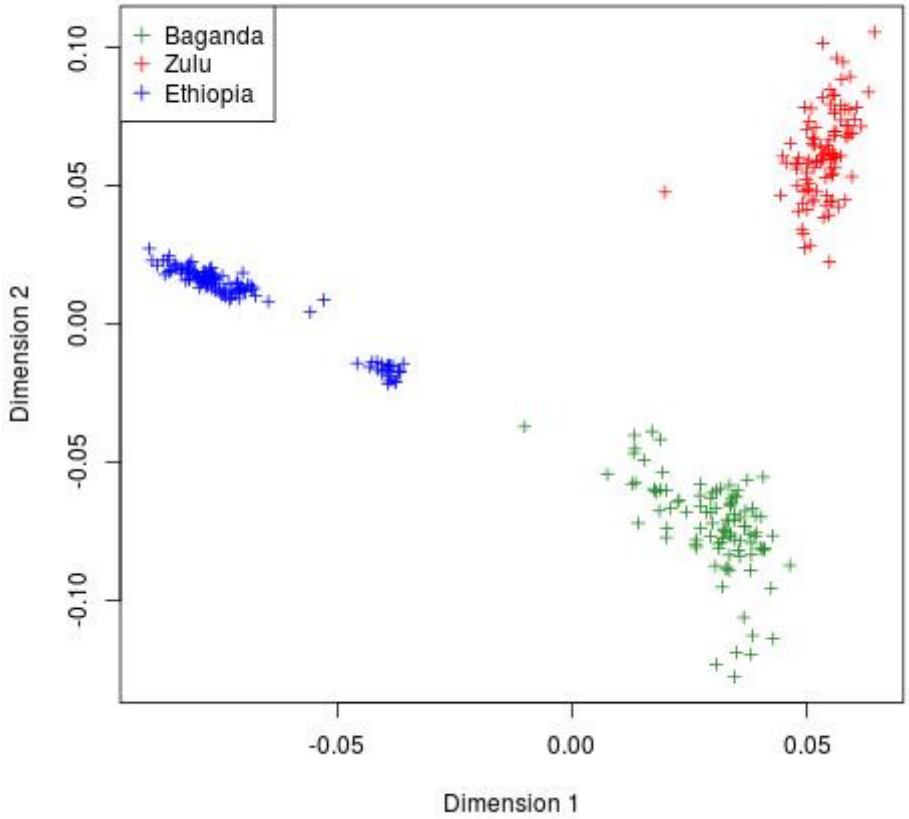
wellcome trust sanger institute

Population structure

- Moderate differentiation between Ethiopians and other populations
- Very little differentiation between Zulu and Baganda

Fst differentiation

	Baganda	Zulu	Ethiopia
Baganda	0	0.008	0.028
Zulu	0.008	0	0.035
Ethiopia	0.028	0.035	0



Generating a new reference panel for imputation

- Several questions about how best to process data:
  - Single population calling vs multiple population calling
  - Can these approaches be combined?
  - Best algorithms for genotype refinement
- Will this panel improve imputation compared to the 1000 Genomes Project panel?

wellcome trust sanger institute

Assessing the utility of extremely low coverage designs

- What coverage is ideal for large scale genomic studies in Africa
- How well can we capture common genetic variation with 2x, 1x and 0.5x coverage data in Africa
- Can we improve capture with a better reference panel
- Trade off of sample size and accuracy

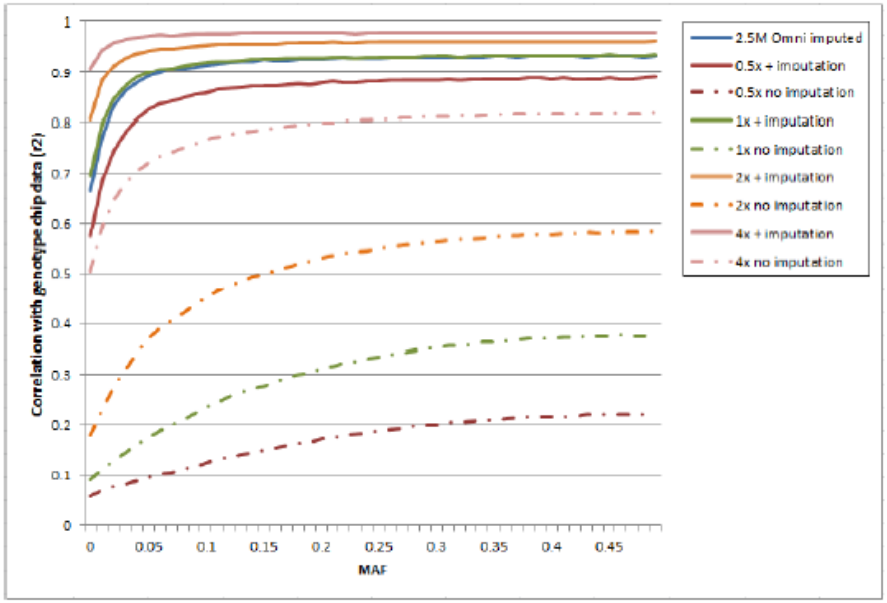
wellcome trust sanger institute

Downsampling genomes

Table 1: A summary of data produced by our pipeline for Ugandan sequence downsampled to different read depths

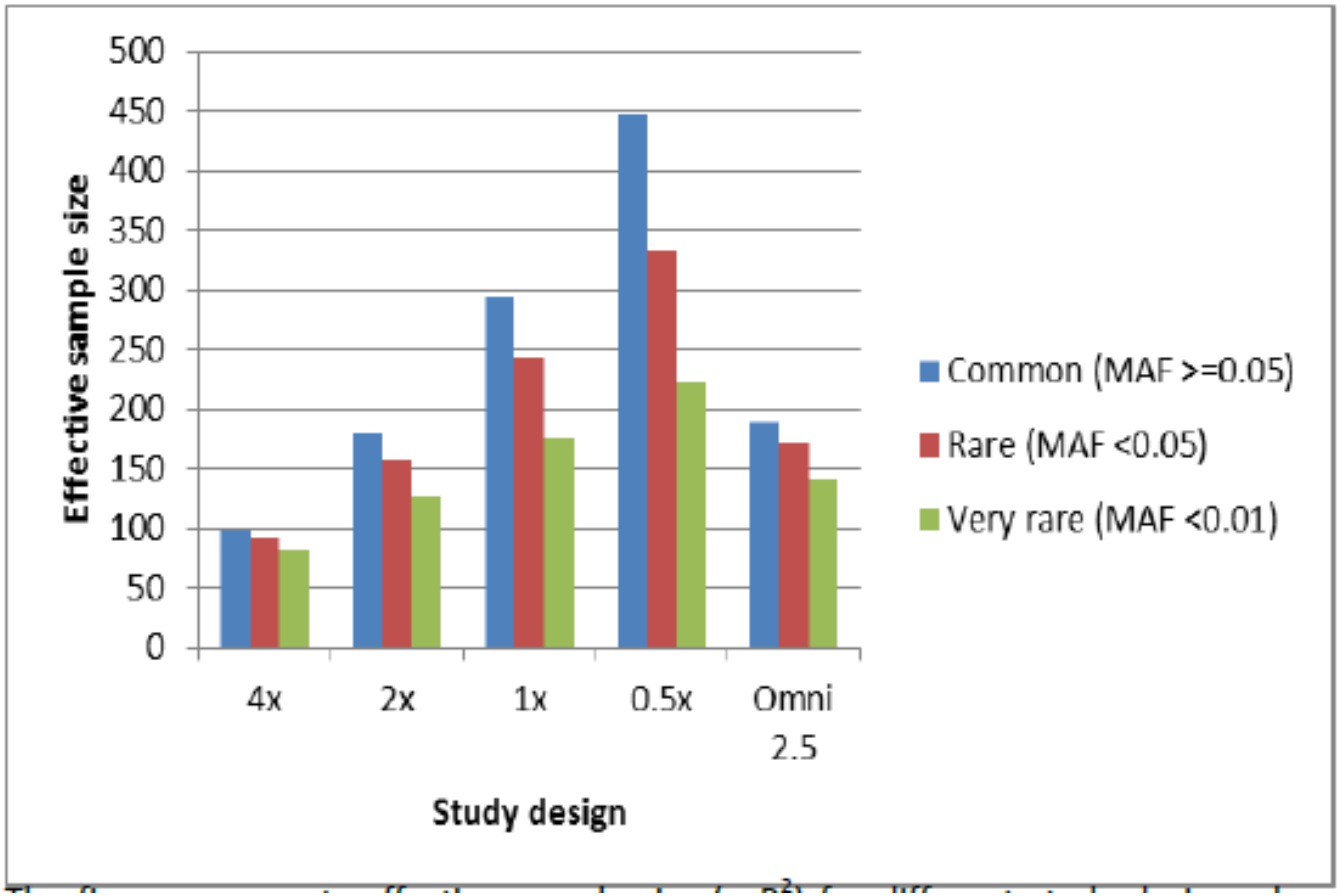
Design	No. of SNPs called	passed filtering			Post-imputation
		Number	dbSNP membership	Ti/Tv Known/Novel	
4x	23,385,101	13,294,444	91.3	2.17/2.08	22,319,833
2x	20,949,992	10,630,167	95.7	2.16/2.00	20,528,334
1x	16,533,092	8,994,440	97.7	2.16/1.92	19,379,195
0.5x	12,750,103	7,522,992	98.9	2.13/1.83	pending

Figure 3: Correlation between called genotype dosages of Ugandan sequence data at different coverages before and after imputation.

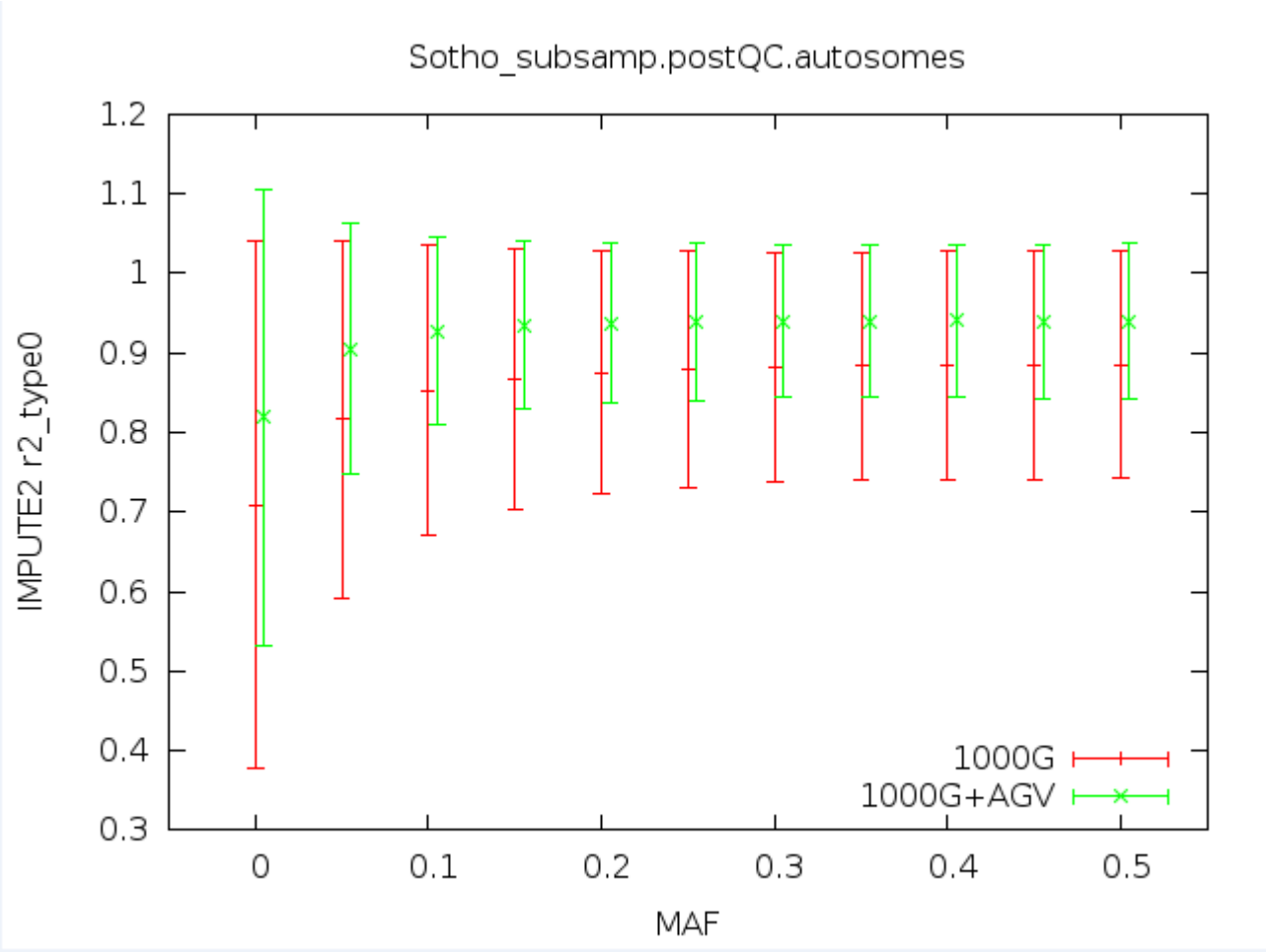


		2x	1x	0.5x	Omni2.5 Imputed
Sensitivity	Common variants (MAF >5%)	0.9840	0.9748	0.9596	0.9563
	Rare variants (MAF<5%)	0.9328	0.9069	0.8884	0.9452
	Singletons	0.6879	0.6033	0.5675	0.7338
	Novel variants †	0.7055	0.6285	pending	0.0000
Specificity	Common variants (MAF >5%)	0.9828	0.9736	0.9559	0.9481
	Rare variants (MAF <5%)	0.9859	0.9865	0.9861	NA
	Singletons	0.9437	0.9527	0.9734	NA
	Novel variants†	0.9617	0.9529	pending	NA

Figure 5: Effective sample sizes for rare and common variants for genome wide association studies for different study designs with a fixed budget



Improving imputation accuracy using a more diverse reference panel



Kyumalibwa cohort study, Uganda

Data collected on more 9,000 participants (Jan to Dec 2011)

—blood pressure, anthropometric measures

—lipids, liver function, FBCs, haemoglobinopathies, hbA1c, renal function, inflammatory and viral response markers

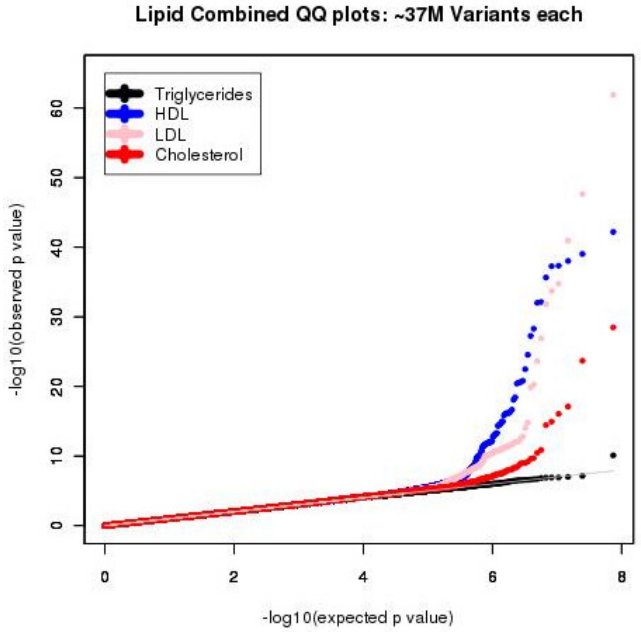
—HIV, HCV, HBV, KSHV

—lifestyle factors and behaviours, medical history

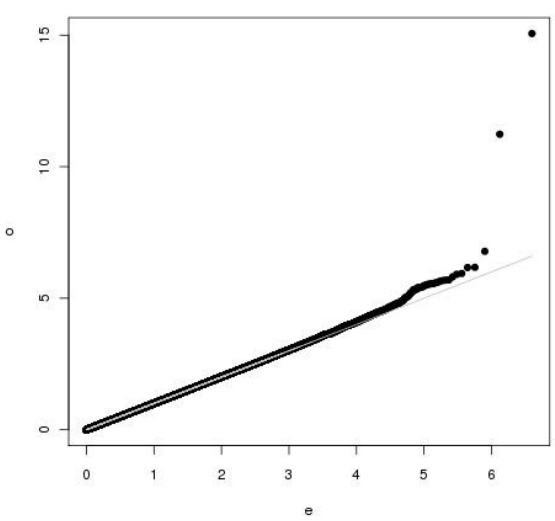
Figure 5. Mapping of village buildings

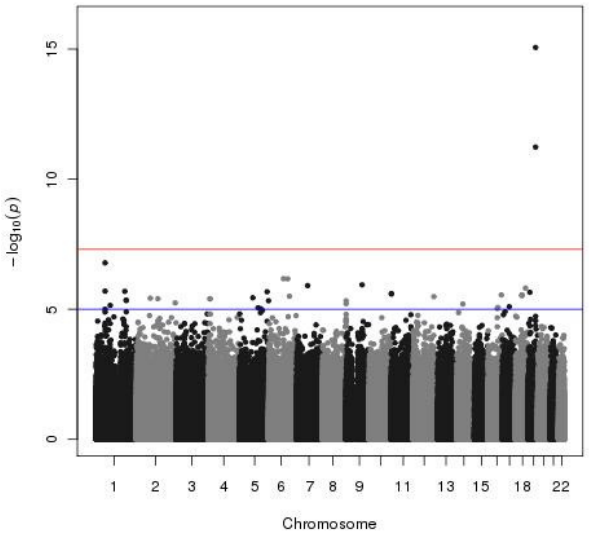
Genomic studies –*Genome wide association studies of cardiometabolic and infectious traits*

- Genotyping of 5,000 participants from Uganda using the 2.5M Illumina with low-coverage whole genome sequencing of a subgroup (100 participants) as an additional imputation reference panel
- Aim to conduct high resolution imputation using family designs
- Provides a framework for whole-genome association studies and global collaboration
- **5,000 samples genotyped so far**
- **100 samples have undergone low coverage whole genome sequencing**



Strategies to conduct genome wide studies in Africa





	Founders only n=419				*All data (including pedigrees) n=2676*			
Algorithm	**SHAPEIT2**	**Beagle**	**SLRP**	**HAPI-UR (3x)**	**SHAPEIT2**	**SLRP**	**Beagle**	**HAPI-UR (3x)**
Switch error rate (SER)*	0.63%	1.4%	0.7%	1.3%	0.12%	0.26%	0.27%	0.31%
SER within SRLP masked segment*	0.30%	0.80%	0.48%	0.81%	0.11%	0.21%	0.21%	0.23%
Yield*	100%	100%	16.8%	100%	100%	68%	100%	100%
Memory	919MB	2.0GB	10.5GB	860MBX3	5.4GB	30GB	4.0GB	3.8GBX3
Run time	18.2h	1.6h	16.9h	0.6hX3	112h	361h	33.9h	6.7hX3

Using GEMMA and PC adjusted to account for relatedness and underlying substructure. Eesults show that this approach works well. **Variance component (linear mixed) model to account for relatedness sample structure in genome-wide association studies**

Phasing strategies for imputation of genetic variants from whole genome sequencing studies

Acknowledgements

WTSI

Deepti Gurdasani
Eleftheria Zeggini
Tommy Carstensen
Savita Karthikeyan
Cristina Pomilla
Georgina Murphy
Elizabeth Young
Ioanna Tachmazidou
Konstantinos Hatzikotoulas

Wellcome Trust

MRC Uganda
Anatoli Kamali
Janet Seely
Pontiano Kaleebu

CRGGH
Charles Rotimi
Daniel Shriner
Fasil Ayele

South Africa

Ayesha Motala
Fraser Pirie
Michele Ramsey
Ananyo Choudhury

Ghana/Kenya/Nigeria
Adebowale Adeyemo
Albert Amoah
Clement Adebamowo
Johnnie Oli

Ethiopia
Chris Tyler Smith
Luca Pagani

Malariagen
Dominic Kwiatkowski

Sanger pipelines

Acknowledgements

WTSI

Deepti Gurdasani
Tommy Carstensen
Elizabeth Young
Cristina Pomilla
Eleftheria Zeggini

MRC Uganda
Anatoli Kamali
Janet Seeley
Pontiano Kaleebu

Oxford
Jonathan Marchini

South Africa

Ayesha Motala
Fraser Pirie
Brenna Henn
Eileen Hoal
Michele Ramsey
Shane Norris

Ethiopia (WTSI)
Chris Tyler Smith
Luca Pagani

Sanger pipeline teams

All participants