Agency for Science, Technology and Research
SINGAPORE
CREATING GROWTH, ENHANCING LIVES

www.a-star.edu.sg/attract
ATTRaCT
Friday October 20, 2017
12:30-1:30pm
2328/F

Genome Institute of Singapore
A*STAR

SG10K: Insights into the genetic architecture of Singaporeans

Bellis C[1], Irwan ID[1], Lin CA[1], Koh TH[1], Wang C[2], Soon WW[3], Wilm A[4], Shih CC[4], & Liu J[1] and the SG10K consortium

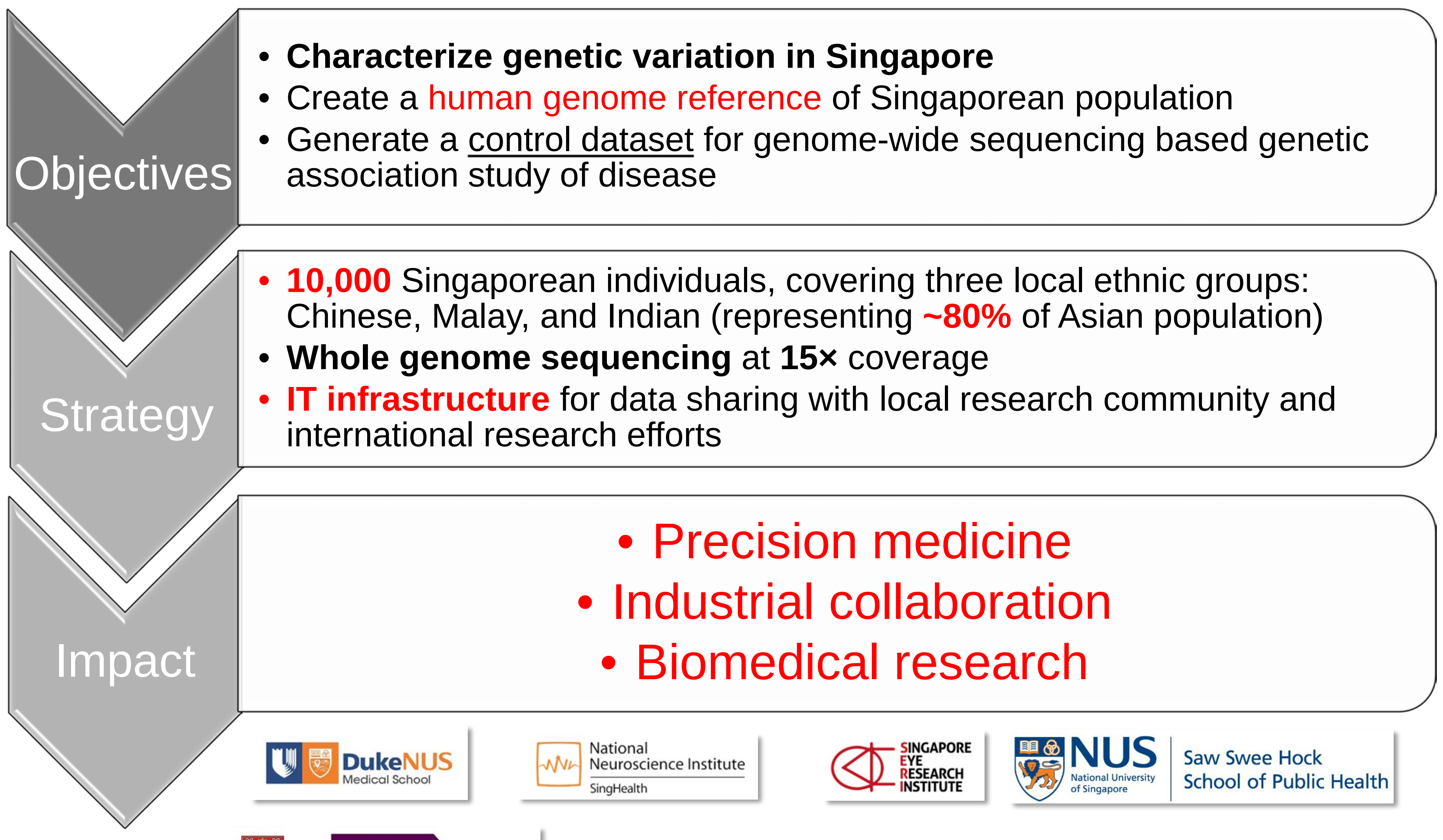
[1]Human Genetics 2, Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR), Singapore • [2]Computational and Systems Biology, Genome Institute of Singapore, A*STAR, Singapore • [3]Next Generation Sequencing Platform, Genome Institute of Singapore, A*STAR, Singapore • [4]Scientific and Research Computing, Genome Institute of Singapore, A*STAR, Singapore

The unique ethnic diversity inherent within the Singaporean population opens it up as an opportune cohort for population genetics studies. The Singapore population consists of three major ethnic groups; Chinese, Malay, and Indian, which together represent ~80% of the genetic variation across Asian populations.

In 2015, the “SG10K project” was initiated with an overarching aim of sequencing the genomes of 10,000 Singaporeans. To date our collaborative partners include SingHealth Duke-NUS Institute of Precision Medicine, Singapore Eye Research Institute, Centre for Personalised and Precision Health, Tan Tock Seng Hospital, National University Health System and several Translational and Clinical Research Flagship Programmes (Heart failure, Parkinson disease).

Our main objectives are to (1) comprehensively characterize genetic variation in Singapore population; (2) create a WGS reference panel for accurate genotype imputation in Asian population; and (3) generate a large control dataset for WGS-based genetic association study of diseases.

We have adopted a shallow-pass sequencing approach, which on average will cover each base at a depth of approximately 15×. Our analytical pipeline hosted by the National Supercomputing Centre (NSCC), Singapore incorporates GATK (v3.6) and follows GATK best practices. Our initial analytical pipeline test was undertaken on n=1,059 genomes and required approximately 2-3 weeks compute time running on 20 reserved nodes at NSCC. Upon completion, this study will provide valuable genetic information to facilitate precision medicine initiatives in Singapore and will empower genetic studies of Singapore and Asian-centric diseases.



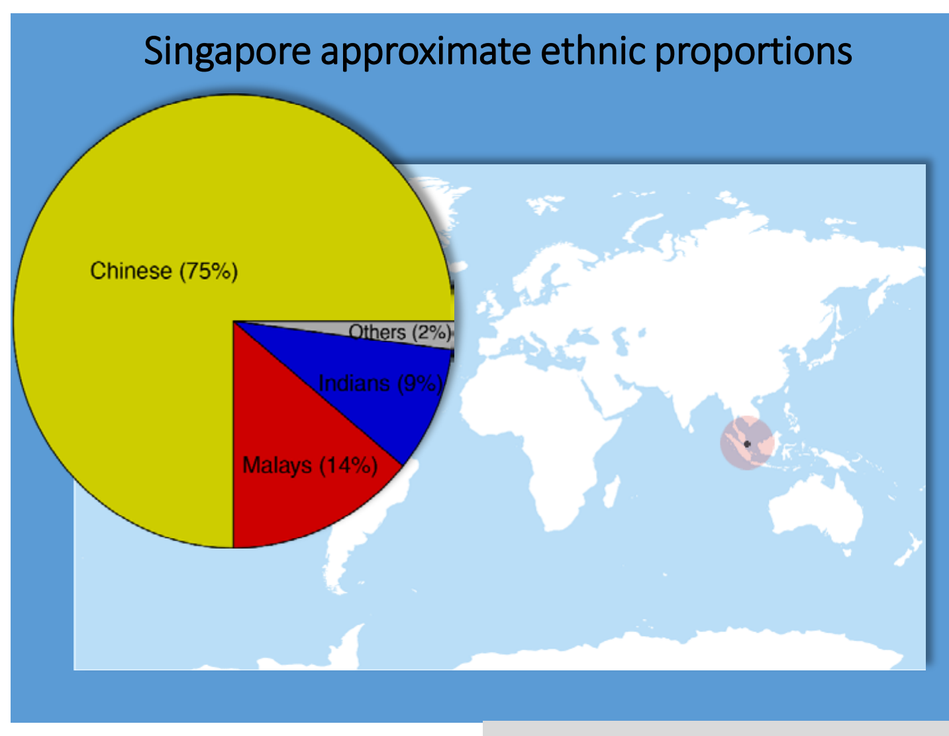
Objectives
- **Characterize genetic variation in Singapore**
- Create a human genome reference of Singaporean population
- Generate a control dataset for genome-wide sequencing based genetic association study of disease

Strategy
- **10,000** Singaporean individuals, covering three local ethnic groups: Chinese, Malay, and Indian (representing **~80%** of Asian population)
- **Whole genome sequencing** at **15×** coverage
- **IT infrastructure** for data sharing with local research community and international research efforts

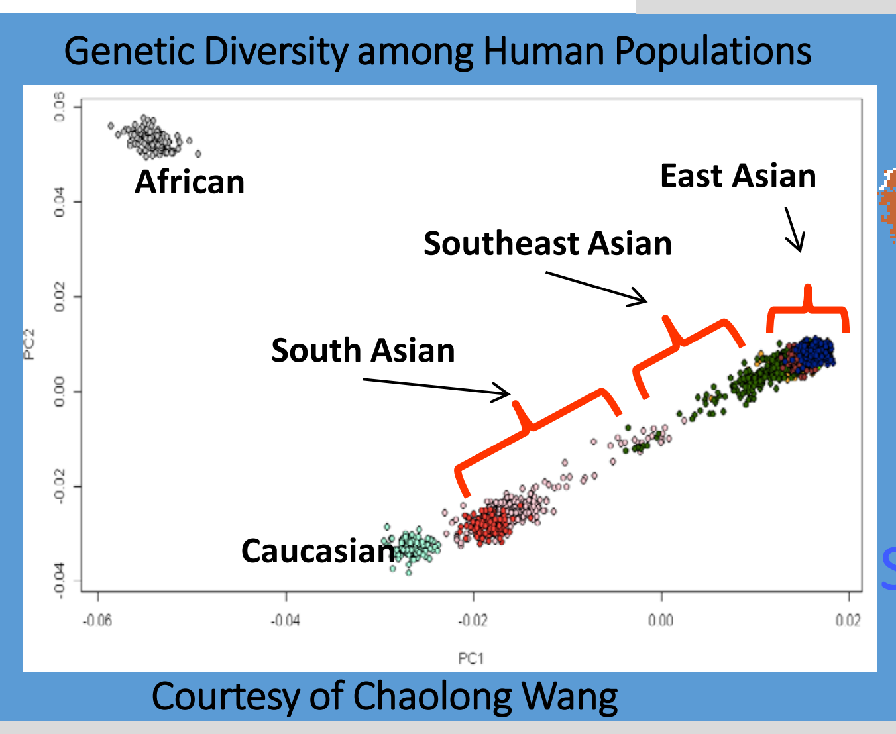
Impact
- Precision medicine
- Industrial collaboration
- Biomedical research

DukeNUS Medical School · National Neuroscience Institute SingHealth · Singapore Eye Research Institute · NUS National University of Singapore Saw Swee Hock School of Public Health · Nanyang Technological University Lee Kong Chian School of Medicine, Imperial College London, Redefining Medicine, Transforming Healthcare · Tan Tock Seng Hospital · NUS National University of Singapore Cardiovascular Research Institute (CVRI)

Context

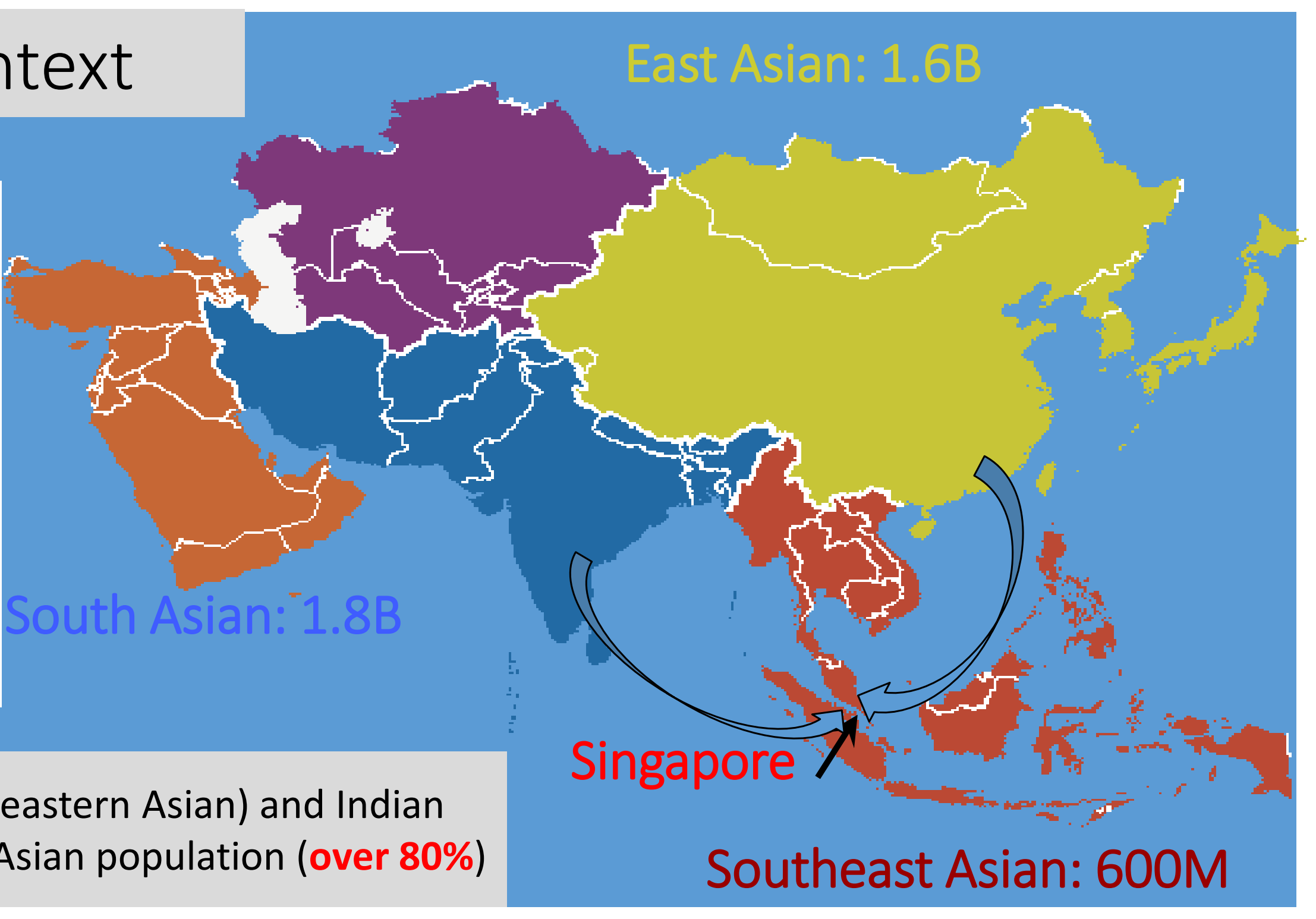


Singapore approximate ethnic proportions



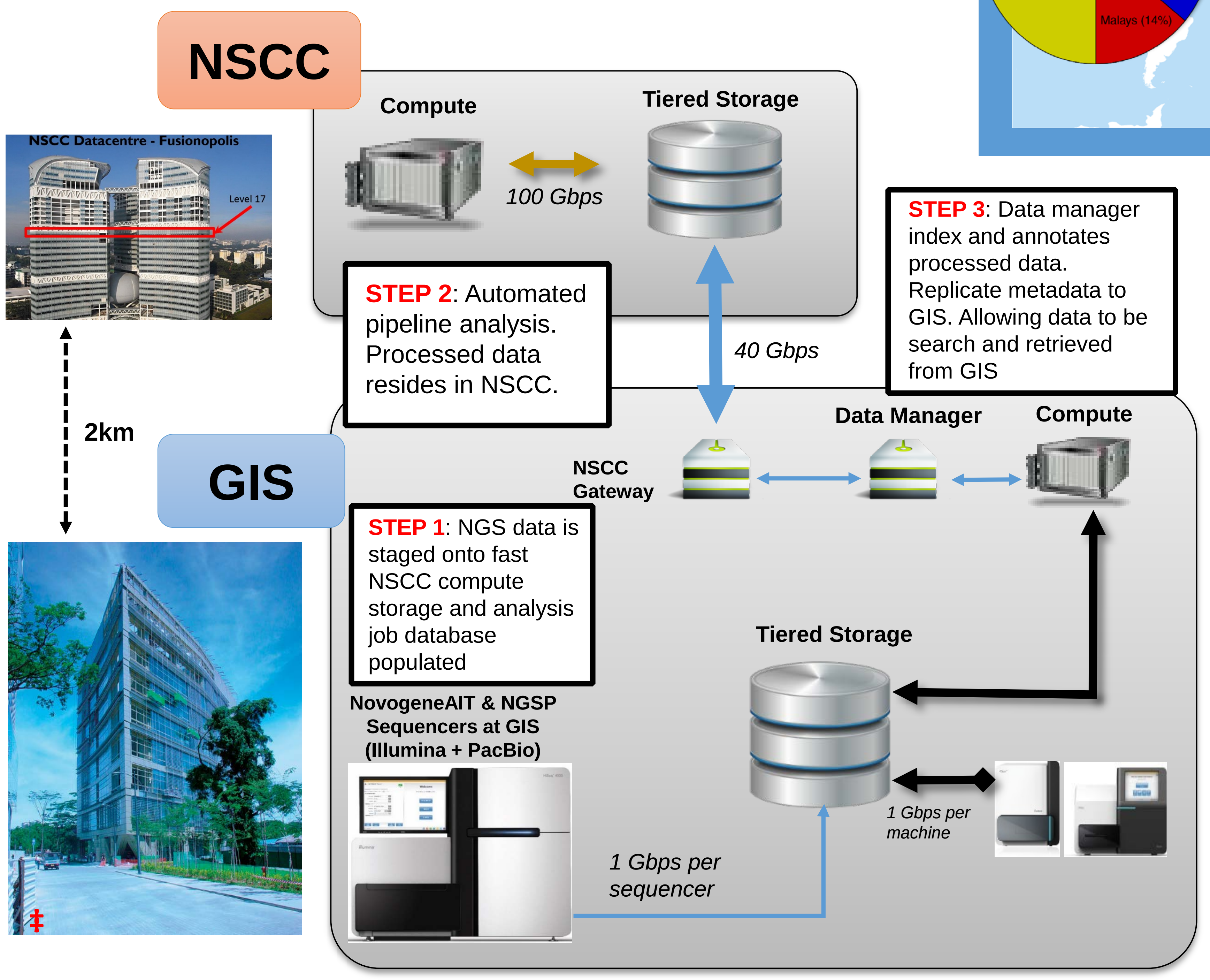
Genetic Diversity among Human Populations

Courtesy of Chaolong Wang



Local Singapore populations of Chinese (East Asian), Malay (Southeastern Asian) and Indian (Southern Asian) gives a good ‘snapshot’ of genetic diversity within Asian population (**over 80%**)

Whole Genome Sequence data: from GIS to remote Supercomputer in NSCC



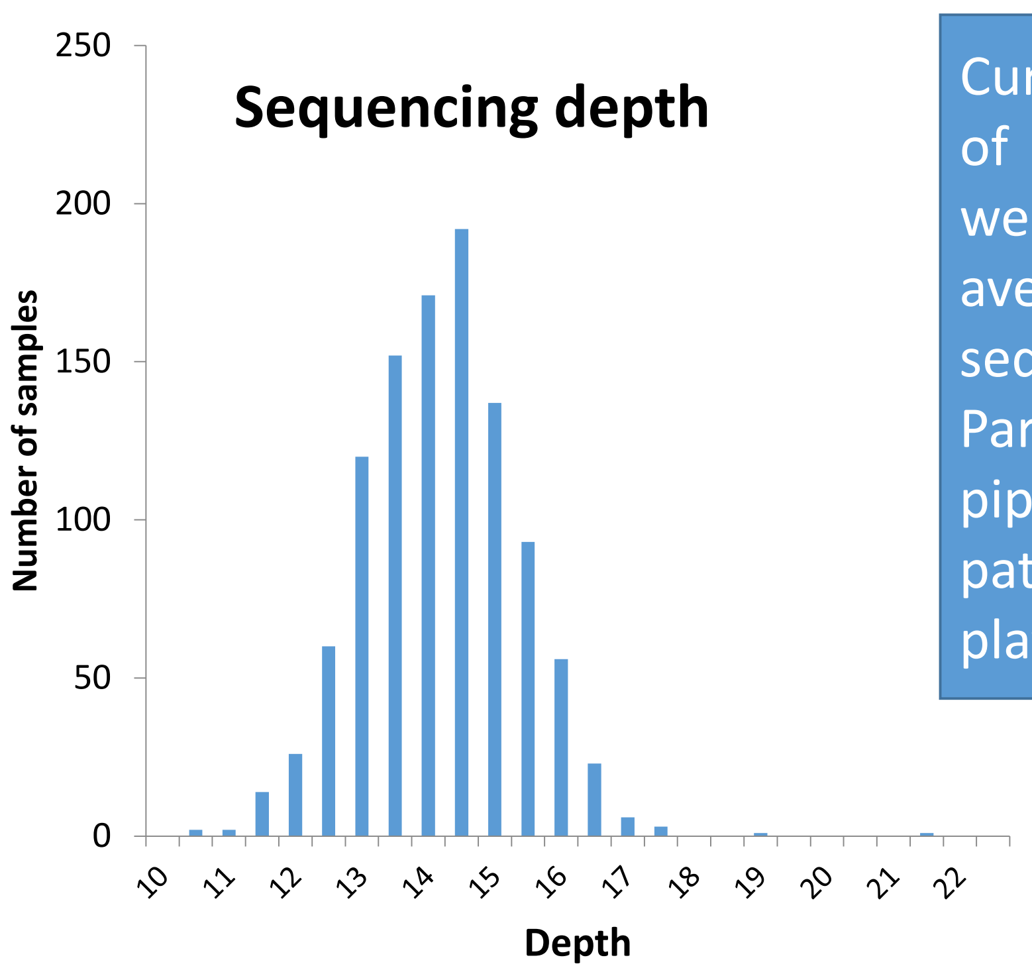
STEP 1: NGS data is staged onto fast NSCC compute storage and analysis job database populated

STEP 2: Automated pipeline analysis. Processed data resides in NSCC.

STEP 3: Data manager index and annotates processed data. Replicate metadata to GIS. Allowing data to be search and retrieved from GIS

GIS Team: JJ Liu (PI), Chaolong Wang (Co-PI), Wendy Soon (Sequencing), Andreas Wilm (Data process pipeline), Shih Chih Chuan (Data storage and database), Claire Bellis (Project Manager)

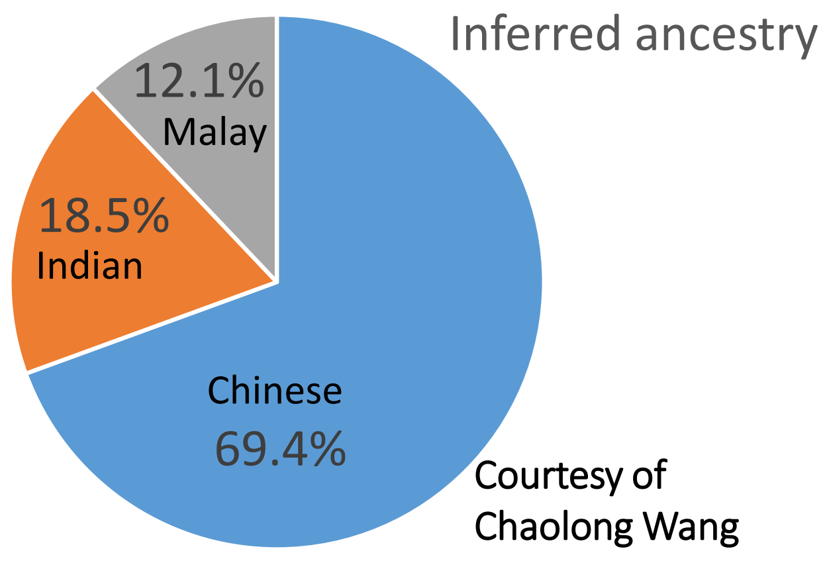
Preliminary Results



Currently, the **SG10K** sequencing project undertaken at Genome Institute of Singapore has sequenced the first n=1,059 genomes. These genomes were used to test and optimize our pipeline for WGS analysis. We report average genome coverage of an expected ~14-15×. Furthermore, we have sequenced a further n=1,500 patient genomes from Heart Failure and Parkinson’s cohorts. These genomes are being processed using the same pipeline as the first 1,000 genomes of SG10K. The analysis of the 1,500 patient samples will allow us to verify the NovogeneAIT sequencing platform as well as quality control and SOP.

Self-reported ancestry appears consistent when compared with estimates inferred by **LASER[1,2]**

	Self-reported Ancestry	Gender % (♂|♀)	Age (♂|♀)
Chinese	64.7%	48|52	54|54
Indian	19.0%	49|51	60|58
Malay	16.3%	49|51	51|49

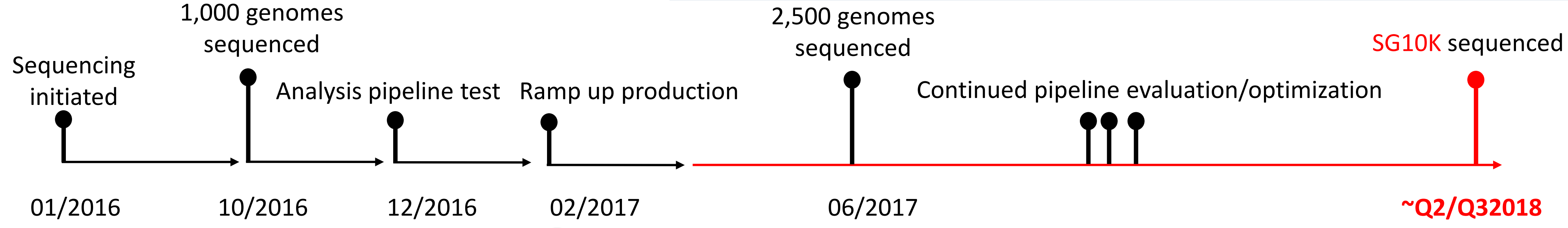


Inferred ancestry

Courtesy of Chaolong Wang

SG10K is well placed heading into the next phase as we aim to sequence our remaining cohorts, totaling n~7,500 samples. We aim to ramp our efforts in the coming months to remain on track with project timelines. The initial pipeline test phase has successfully completed with preliminary metrics available for interpretation.

Projected timeline



ASHG 2017 ORLANDO • OCTOBER 17-21, 2017
SHARING DISCOVERIES. SHAPING OUR FUTURE.

[1]Wang C *et al.,* (2014) *Nat Genet.* doi:10.1038/ng.2924
[2]Wang C *et al.,* (2015) *AJHG* dx.doi.org/10.1016/j.ajhg.2015.04.018

‡ https://commons.wikimedia.org/w/index.php?curid=1370607