Presentation title: Population-scale high-thoughput sequencing data analysis

Abstract:
Unprecedented computational capabilities and high-throughput data collection methods promise a new era of personalised, evidence-based healthcare, utilising individual genomic profiles to tailor health management as demonstrated by recent successes in rare genetic disorders or stratified cancer treatments. However, processing genomic information at a scale relevant for the health-system remains challenging due to high demands on data reproducibility and data provenance. Furthermore, the necessary computational requirements requires a large investment associated with computer hardware and IT personnel, which is a barrier to entry for small laboratories and difficult to maintain at peak times for larger institutes. This hampers the creation of time-reliable production informatics environments for clinical genomics. Commercial cloud computing frameworks, like Amazon Web Services (AWS) provide an economical alternative to in-house compute clusters as they allow outsourcing of computation to third-party providers, while retaining the software and compute flexibility.

To cater for this resource-hungry, fast pace yet sensitive environment of personalized medicine, we developed NGSANE, a Linux-based, HPC-enabled framework that minimises overhead for set up and processing of new projects yet maintains full flexibility of custom scripting and data provenance when processing raw sequencing data either on a local cluster or Amazon's Elastic Compute Cloud (EC2).