*Email*

Please come to the interview ready to discuss the challenges we face supporting scientists providing data to the Human Cell Atlas (HCA).

An important challenge facing our team is ensuring all the collected data is well structured and described. This is vital to enable the definition of the atlas and enable scientific discovery using the resource.

The groups providing data to the HCA will range from large centres like the Sanger and the Broad to small labs who operate using Excel spreadsheets.

During the interview, you will need to describe the solutions you would propose the group pursues to solve these problems.

*HCA key players:*

Organizing Committee is co-chaired by **Aviv Regev (Broad)** and **Sarah Teichmann (Wellcome)**

Other organizing committee members:

Michael Stratton (Wellcome)

Peter Campbell (Wellcome)

Nikolaus Rajewsky (Germany)

Eric Lander (Broad)

Jonathan Weissman (UCSF)

*Panel Interview:*

**Laura Clarke** leads resequencing informatics in the Vertebrate Genomics group at EMBL-EBI. Resequencing Informatics provides data management, primary analysis, and data distribution services for large sequencing consortia like the 1000 Genomes Project, Blueprint, HipSci and FAANG.

**John Marioni**’s group develops the computational and statistical tools necessary to exploit high-throughput genomics data, with the aim of understanding the regulation of gene expression and modeling developmental and evolutionary processes.

**Thomas Keane**’s group is interested in using genomic technologies to learn about biological processes, with a particular focus on mouse and human disease.

**Robert Petryszak** leads the Gene Expression team, which deals with the acquisition, curation, quality control, statistical analysis, and visualisation of functional genomics data, as well as user support and training for Expression Atlas.

***How can we ensure that all the collected data is well structured and described?***

1. Strict requirement of **data** allowed to be submitted
   1. Raw
      1. Demultliplexed NGS runs (fastq)
      2. FISH images (?)
      3. Others?
   2. Processed data
      1. Derived analyses and quality metrics from running vetted analysis pipelines
      2. Could be problematic, lose provenance if workflows aren’t submitted
   3. Raw data MUST be accompanied by compete metadata at time of submission
2. Strict requirement of complete **metadata** submission at the time of data upload.
   1. Two methods for submission
      1. Manually enter metadata into **web-interface**
         1. For smaller labs generating fewer datasets by providing a simple web interface
         2. Allow for embedded chat (gitter, etc.)
         3. Viewing previous metadata/sample uploads
         4. Use drop-down and select from list options, preventing users from making inappropriate choices. Less “free text” fields. Galaxy implements this in conversion tools.
      2. Generate **metadata file** and upload to data broker
         1. For larger institutes (Sanger/Broad) that are generating many datasets by providing or helping to provide scripts that automatically generate a metadata file (yaml, json, etc) compatible with the HCA data broker
         2. Need to be pre-processed before upload to ensure fields adhere to HCA standards
   2. Metadata itself
      1. No Excel files! GEO does not do a good job of this: uses Excel files which are classist (not every lab can purchase MS license) and lossy (conversion to date format).
      2. Editing of metadata (and data) needs to be highly controlled, with a ticketing system and version control
      3. Experimental protocols needs to be standardized using something like protocols.io and should be easily displayed at HCA
3. Submission of processed data invites some issues
   1. Although processed data would be generated using “community-agreed upon workflows”, what does this mean?
   2. Need to ensure tool versioning and parameter choice control (CWL?)

1. Ingestion Service API - is this already developed?

1. Make available published workflows from other groups (if they wish), links to publications using the HCA data

1.