



Practical – Genomics England Embassy

Getting at the Data

Matthew Parker

Lead Bioinformatician

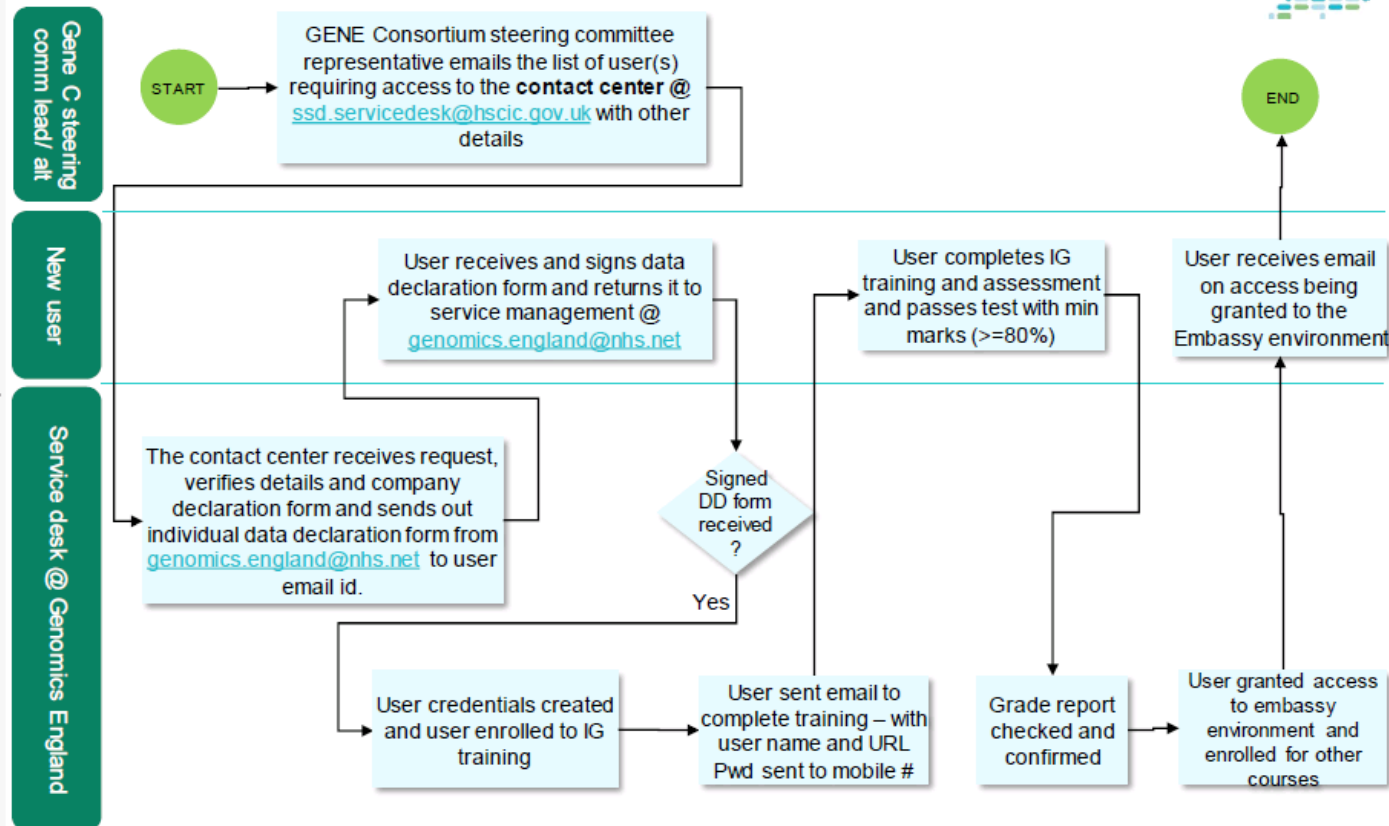
Sheffield Diagnostic Genetics Service

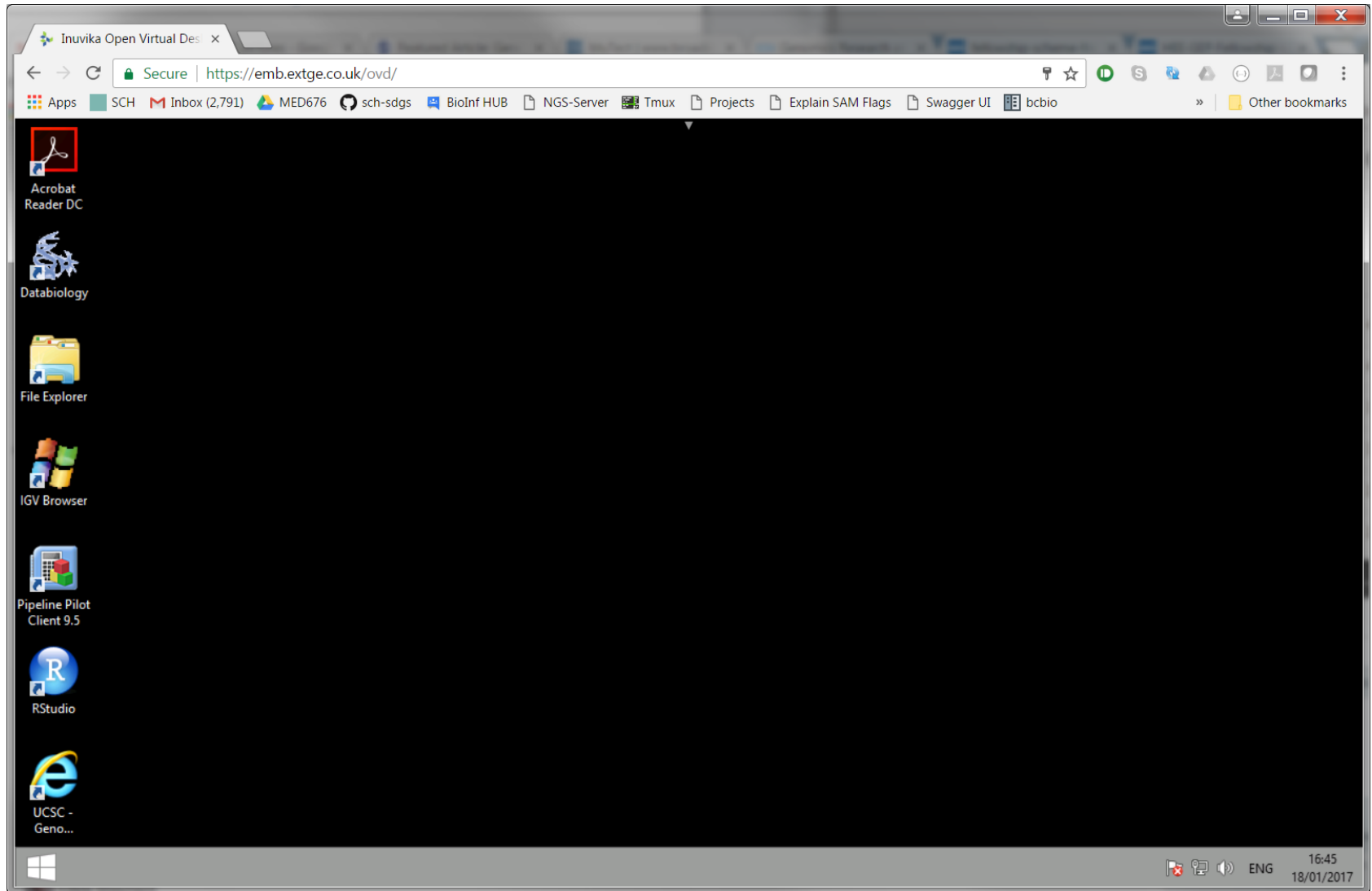
Sheffield Children's **NHS**
NHS Foundation Trust

- Patient confidentiality and protection is a key cornerstone of the 100,000 Genomes Project.
- individual level data will not be “released” but will instead be analysed within a secure, monitored environment akin to a reading library
- To protect patient confidentiality access to this environment will be reviewed and granted only for specific, approved purposes in accordance with informed consent and the ethical scope of the Protocol, and any attempted usage beyond the specified purpose may lead to exclusion and possible legal action, supported by automatically-generated electronic evidence.

- Secure area to query GEL repository, store results etc
- Browser based virtual desktop interface
- Access controlled by GEL
- Not everyone has access to the embassy: we'll talk through how to get access and the structure of the data
- The education embassy is currently very limited – not many tools or much data is available

New user Access request process





<https://emb.extge.co.uk/ovd/>

Structure of the Data

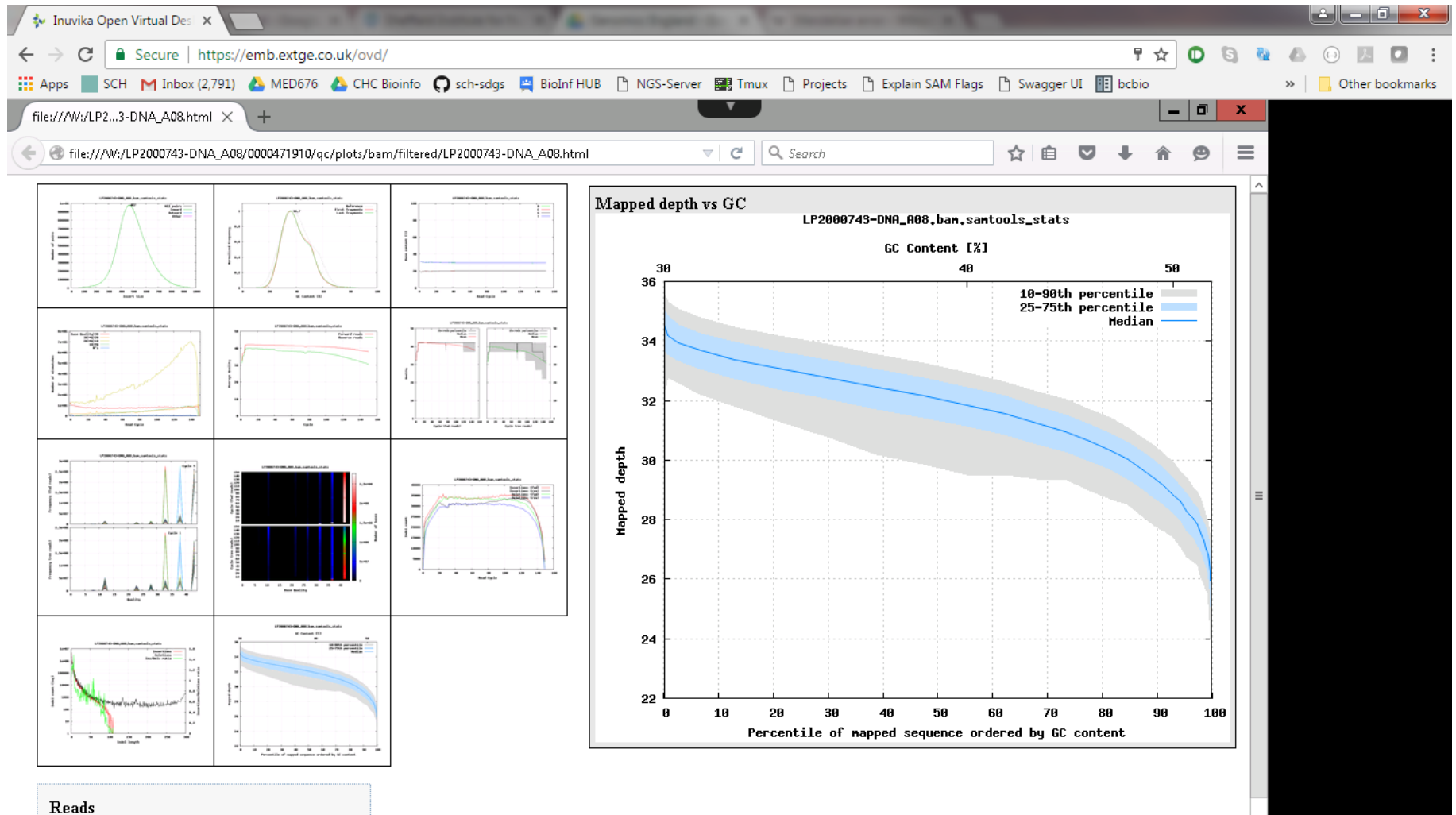
The image shows a virtual desktop environment with two windows. The top window is a web browser displaying the URL <https://emb.extge.co.uk/ovd/>. The bottom window is a terminal window showing the output of the `tree` command for the directory `edu_gecip/LP2000739-DNA_F02/`.

```
dwang@EMB-LINUX02:~$ tree edu_gecip/LP2000739-DNA_F02/
edu_gecip/LP2000739-DNA_F02/
├── 0000467102
│   ├── Assembly
│   │   ├── LP2000739-DNA_F02.bam
│   │   └── LP2000739-DNA_F02.bam.bai
│   ├── Genotyping
│   │   ├── LP2000739-DNA_F02.GenotypingReport.txt
│   │   ├── LP2000739-DNA_F02.Genotyping.vcf.gz
│   │   ├── LP2000739-DNA_F02.Genotyping.vcf.gz.tbi
│   │   ├── LP2000739-DNA_F02.idats
│   │   ├── 3999062073_R04C01_GRN.idat
│   │   ├── 3999062073_R04C01.gtc
│   │   ├── 3999062073_R04C01_RED.idat
│   │   └── WG0322059MSA3_sampleSheet.csv
│   ├── LP2000739-DNA_F02.SummaryReport.csv
│   ├── LP2000739-DNA_F02.SummaryReport.pdf
│   ├── md5sum.txt
│   ├── Metrics
│   │   ├── LP2000739-DNA_F02.baseCompositionPerCycle.csv
│   │   ├── LP2000739-DNA_F02.GCDistribution.csv
│   │   ├── LP2000739-DNA_F02.insertSizeHistogram.csv
│   │   ├── LP2000739-DNA_F02.Metrics.csv
│   │   ├── LP2000739-DNA_F02.Qscore_mean_byCycle.csv
│   │   └── LP2000739-DNA_F02.uniformityOfCoverage.csv
│   └── qc
│       ├── logs
│       │   ├── DAPI_reporting.log
│       │   ├── LP2000739-DNA_F02_CATALOG_REGISTRATION_20150925-051842.log
│       │   ├── LP2000739-DNA_F02_CATALOG_REGISTRATION_20150929-095512.log
│       │   ├── LP2000739-DNA_F02_CATALOG_REGISTRATION_20150930-025926.log
│       │   ├── LP2000739-DNA_F02.metrics_bam.log
│       │   ├── LP2000739-DNA_F02.validate_bam.log
│       │   ├── LP2000739-DNA_F02.validate_vcf.log
│       │   ├── plot-vcf-LP2000739-DNA_F02.genome.vcf.gz.log
│       │   ├── plot-vcf-LP2000739-DNA_F02.SV.vcf.gz.log
│       │   └── plot-vcf-LP2000739-DNA_F02.vcf.gz.log
│       ├── LP2000739-DNA_F02.bam.metrics
│       ├── LP2000739-DNA_F02.bam.picard_validateSAMFile
│       ├── LP2000739-DNA_F02.bam.Q30
│       └── LP2000739-DNA_F02.genome.vcf.gz.bcftools_warning
```

The right window is a file explorer showing the directory structure of `edu_gecip (W:) > LP2000739-DNA_F02 > 0000467102`. The search bar contains `Search 0000467102`.

Name	Date modified	Type	Size
Assembly	25/06/2015 17:25	File folder	
Genotyping	19/06/2015 03:44	File folder	
Metrics	19/06/2015 03:44	File folder	
qc	03/07/2015 15:19	File folder	
Variations	19/06/2015 03:48	File folder	
LP2000739-DNA_F02.SummaryReport	18/06/2015 21:52	OpenOffice.org 1...	3 KB
LP2000739-DNA_F02.SummaryReport	18/06/2015 21:52	Adobe Acrobat D...	125 KB
md5sum	18/06/2015 21:52	Text Document	2 KB

- Samtools – tool for interrogating sequence files – also includes ability to generate stats on these files
- W:\LP2000743-DNA_A08\0000471910\qc\plots\bam\filtered\LPS000743-DNA_A08
- Low level qc of the sequencing run



IGV Viewing BAM & VCF

