Andrew J Skelton

☎ 07890931710 ✉ Andrew.J.Skelton73@gmail.com

Newcastle Upon Tyne, UK
github.com/AndrewSkelton
andrewskelton.github.io
@ajskelton73

Data Scientist and Bioinformatician working with Clinical and Genomic Data

MSc (Hons) Bioinformatics
– Distinction, Prize Winner
BSc (Hons) Computer Science

Current Role:

Senior Bioinformatician, Newcastle University Medical School - 3.5 Years

Current Languages:

Day to Day: *R*	Cloud: *AWS*
Text processing: *R/Python*	Obj Orientated: *Java*
Web Framework: *R-Shiny*	Database: *Mongodb*

Publications

Science Translational Medicine: Human IFNAR2 Deficiency: Lessons for Antiviral Immunity *doi:10.1126/scitranslmed.aac4227*

The Journal of Biological Chemistry: Cytokine-Induced MMP13 Expression in Human Chondrocytes Is Dependent on Activating Transcription Factor 3 (ATF3) Regulation *doi:10.1074/jbc.M116.756601*

PloS One: Leptin and Pro-Inflammatory Stimuli Synergistically Upregulate MMP-1 and MMP-3 Secretion in Human Gingival Fibroblasts *doi:10.1371/journal.pone.0148024*

BMC Medical Genetics: Expression Analysis of the Osteoarthritis Genetic Susceptibility Locus Mapping to an Intron of the MCF2L Gene and Marked by the Polymorphism Rs11842874 *doi:10.1186/s12881-015-0254-2*

+ *1 First Author publication in progress, and 3 publications pending*

Conference Talks

The First International Workshop on the Epigenetics of Osteoarthritis: Analysis of Illumina Human Methylation Array Data using linear models *doi:10.3109/03008207.2016.1168409*

Employment History

2012	2013	Jan 2014 - Dec 2015	Jan2016-Present
Huntsman Corporation	Huntsman Corporation	Newcastle University Medical School	Newcastle University Medical School
Database Design	System Automation	Bioinformatics Analyst	Senior Bioinformatician

Complete BSc Computing Science

Complete MSc Bioinformatics

Current Role

Senior Bioinformatician

Newcastle University Medical School
Jan 2016 - Present

I've been employed by Newcastle University Medical School as a Bioinfomatician since January 2014, under the banner of the Bioinformatics Support Unit. I provide support, expert analysis on high throughput genetic experiments, and general data science problems. I work with multiple groups across two institutions; the institute of genetic medicine (IGM), and the institute of cellular medicine (ICM), mainly encompassing immunology and muskuloskeletal research. I provide support to Professors, PIs, PhD, and Masters students, as well as external collaborators. As bioinformatics is a specialism, I've also provided training to academics on command line linux, and R.

Routine Tasks

- Lead Analyst on multiple projects simultaneously.
- Consult on experimental design, feasibility, and costing of high throughput experiments.
- Apply appropriate statistical frameworks for hypothesis testing up to the scale of 10^{12} tests.
- Design and implement analyses to be highly efficient, leveraging petabyte scale HPC systems where necessary, and parallelising.
- Provide R / Bioinformatics training to a range of staff.
- Familiarity with modern bioinformatics tools, databases, and software in the context of human medical research.
- Supervision of PhD student projects, and masters projects.
- System administration of Linux servers and HPC systems.
- Mentoring junior staff and their personal development.
- Implementation of storage solutions for managing sequencing data.
- Consultation for factory scale sequencing solution, and respective compute / storage.

Skills

Bioinformatics

Lead analyst on several bioinformatics analyses, including RNA seq, Exome Seq, array based assays, epigenetic arrays, and others. Designed and implemented bespoke analyses tailored around disease specific hypotheses, and statistically robust models around experiments. Familiar with packages such as; DESeq2, Limma, for differential expression modelling, STAR, HISAT2, Salmon, Kallisto, for abundance estimation. Worked with public databases and repositories of biological data such as ArrayExpress, GEO, Ensembl, UCSC, and data integration problems. Analysis strategies around rare diseases.

- Traditional alignment methods (BWA, bowtie2, STAR, HISAT2)
- Variant analysis pipelines, including familiarity with GATK's best practices.
- Differential expression modelling (Limma, edgeR, DESeq2)
- Complex experimental design accounting for confounding effects and variables
- Dataset combination and evaluation where feasible.
- Array technologies (Gene Expression, Illumina HumanMethylation, Genotype chips).
- Familiarity with genome assemblies, and differences between annotation methodologies.
- Alignment-free methods (Kallisto, Salmon, Sleuth).

Data Science

Comfortable in Linux environments, and experience in system administration of Debian servers. Regularly refactor code to run on HPC architecture through grid engine. Machine learning for predictive classification problems using Caret and e1071. Exploring the use of methylation arrays for disease classification, however large feature sets (~850,000), pose some challenges for biological data.

- Applying methods to large public datasets.
- Assess analytic performance, improve speed, and memory efficiency.
- Design and implement bespoke analyses around new frameworks.
- Distributed dataset analysis leveraging RHadoop, and Apache Spark (Sparklyr).
- Predictive classification model development for medical applications
- Forecasting using new tools such as Prophet.
- Interpret complex results and present for non-expert audience.

Statistics

Strong foundation of statistics from masters degree, courses and current role. Experience in applying complex statistical methods to large datasets to test specific hypotheses. Often consulted and advised on appropriate statistical tests for large projects. Applying statistical methods in R efficiently, maximising its vectorised implementation.

- Visualisation of results, in simple and effective ways.
- Model design, fit, and evaluation
- Automation of reporting.
- Power and sample size analyses around datasets for estimating experimental parameters.
- Benchmarking performance of new methods

Programming

Day-to-day programming in R and highly proficient in several key language specific constructs including; RMarkdown (Rmd), the Shiny web framework, ggplot2, and the "tidyverse" suite of packages. Familiar with Python for text processing and some web frameworks. Recently started exploring Scala as a progression from Java. Developed robust pipelines designed to scale on HPC resources using grid engine, in bash.

- R, RStudio, Rmd, Shiny, ggplot2, (d)plyr, tidyverse.
- Object orientated programming concepts, Java.
- Bash pipeline design.
- Some Python experience for specific text processing purposes.
- Experienced in pipeline design/ implementation around Bash.
- SQL and NoSQL databases.
- WDL / CWL

Project of Interest

PID Diagnostic System

/PID-WES-GATK3.4-SGE and /Exome-Utilities

Primary immunodeficiency (PID) diagnostic project. Patients routinely come to the Great North Children's Hospital for specialist diagnoses in relation to immunodeficiency, often standard panels fail to identify complex underlying genetic causes, due to causal rarity. Patient and family member DNA then exome sequenced. I designed and implemented a system that takes raw data and produces a set of high quality variant calls, alongside structural candidates. Robust to differences in chemistry, sequencing instrument, and, cross-batch pedigrees, in which effects are absorbed. Refined call-sets are annotated and loaded into a web framework which allows for complex Mendelian inheritance queries in a fast, friendly manor for specialised doctors and researchers to assess. The project is continuous, growing by 40GB (raw compressed data) per month.

- 18 Subprocesses, elegantly scalably to HPC architecture, with 6 core procedures.
- 2 days computation from recieving sample to potential diagnosis.

>200 Samples

35 Families

1.5TB Compressed
Raw Sequencing Data