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

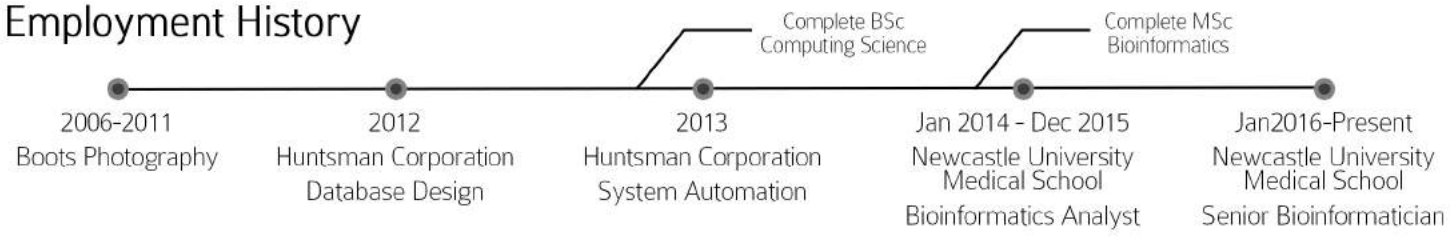
BMC Medical Genetics: Expression Analysis of the Osteoarthritis Genetic Susceptibility Locus Mapping to an Intron of the MCF2L Gene and Marked by the Polymorphism Rs11842874 *doi:10.1186/s12881-015-0254-2*

+ 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



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

I've worked on a variety of projects being the lead analyst on several bioinformatics problems, including RNA seq, Exome Seq, array based assays, epigenetic arrays, etc. I've written numerous bespoke analyses tailored around disease specific hypotheses, and designing robust models around experiments. I'm familiar with packages such as; DESeq2, Limma, for differential expression modelling, STAR, HISAT2, Salmon, Kallisto, for quantification and analysis tasks. I've worked with public databases and repositories of biological data such as ArrayExpress, GEO, Ensembl, UCSC, and data integration problems, such as gene expression and methylation inverse correlations.

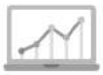
- Traditional alignment methods (BWA, bowtie2, STAR, HISAT2)
- Variant analysis pipelines, including familiarity with GATK's best practices.
- Differential expression modelling (Limma, edgeR, DESeq2)
- Complex experimental design accounting for confounding effects and variables
- Dataset combination 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

I'm comfortable in Linux environments, and have experience in system administration of Debian servers. I also regularly refactor code to run on HPC architecture through grid engine. Machine learning has been a topic of interest to me in the bioinformatics field, namely guaussian mixture models, and random forest techniques in detecting true positive variants in human genomic data, as such I've become familiar with some packages such as Caret, and e1071. I've recently been exploring the possibility of using methylation assays for disease classification, however with a large feature set (~850,000), this poses 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



Statistics is the cornerstone to successful data analysis, and in the context of medical research, the correct statistical framework can drive research forward quickly and effectively. I have a strong foundation of statistics from my masters degree, and I've strived to build these skills while in my current role. I have experience in applying complex statistical methods to large datasets to test specific hypotheses. I've often consulted and advised on appropriate statistical tests for varying data types on large projects which has led to a better understanding of underlying results. Additionally, I have extensive experience applying statistical methods in R in efficient ways, maximising its vectorised implementation.

- Interpret complex results and present for non-expert audience.
- Assess analytic performance, improve speed, and memory efficiency.
- Design and implement bespoke analyses around new frameworks.
- 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.

Programming

The R statistical programming language is my day-to-day language, as such I've become highly proficient in several key language specific constructs including; RMarkdown (Rmd), the Shiny web framework, ggplot2, and the "tidyverse" suite of packages. I'm well acquainted with dplyr's grammatical methodology of dataset cleaning, and applying many statistical methodologies to datasets, where appropriate. I'm familiar with Python for text processing and some web frameworks. I've recently started exploring Scala as a natural progression from originally learning Java during my undergraduate degree, and working in big data. Working with large scale genomic data, I've developed robust pipelines designed to scale on HPC resources using grid engine.

- R, RStudio, Rmd, Shiny, ggplot2, (d)plyr, tidyverse.
- Understanding of object orientated programming concepts, specifically Java.
- Some Python experience for specific text processing purposes.
- Experienced in pipeline design/ implementation around Bash.

Projects of Interest

PID Diagnostic System

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

A project I've worked on for an extended period is the primary immunodeficiency (PID) diagnostic project. Patients routinely come to the Great North Children's Hospital for specialist diagnoses in relation to immunodeficiency, and often standard panels fail to identify complex underlying genetic causes. Patient DNA is taken, often family members in many cases, and exome sequenced. I designed and implemented a system that takes sequenced samples and produces a set of high quality variant calls, alongside structural candidates. The system is robust to differences in chemistry, sequencing instrument, and, cross-batch pedigrees, in which effects are absorbed. Post variant calling, the call-sets are quality assessed, and annotated before being 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

Osteoarthritis Methylation International Collaboration

I developed a protocol for the analysis of Illumina Infinium 450K Methylation Arrays, broken down into analysis modules. This protocol was used by institutes including Oklahoma and Liden to process their Osteoarthritic samples. I designed analyses to identify CpGs related to disease stratification, sites that correlate with age, and probes that show non-linear relationships to age. Additionally, I've been developing an analysis that associates differentially methylated CpGs with potential transcription factor binding sites, to identify enriched transcription factors. Further to this, I've expanded the protocol to allow the combination of assays from various projects, leveraging control probes to estimate and regress technical variance.

NMF Masters Project (*Distinction*)

I worked with the Paediatric Leukaemia Cytogenetic Research group, based in the Royal Victoria Infirmary for my final research project during my master's degree. I implemented a distributed, and parallelised data dimensionality reduction algorithm. I developed a novel variation of NMF – Non-Negative Matrix Factorisation, which calculated an optimum value of *r*, or the number of 'clusters' in an input matrix *m*. This allowed the inference of potential underlying interactions in gene expression datasets.

Volunteering

Athletic Union (AU) Committee Member

I was elected as a member of the AU committee during my Masters year, in which I was responsible (along with 4 other elected members), for club discipline actions, shaping AU policies, and large events organisation. This was a university wide committee, as such it was a lot of responsibility and commitment.

Offensive Line Coach - Newcastle University American Football (NU Raiders)

During my masters degree, I was president of the American Football club at Newcastle University, after 3 years as a player. For my contribution as president, I was awarded Club Colours, and for reaching the national final I was awarded Half-Colours. Post graduation, I vonunteered as a coach, specifically the offensive line position. I coach 2-3 times a week, and find it extremely rewarding seeing player development.