**Statistical Genomics Workshop**

**Lecturers:**

TPS: Terence Paul Speed (UC Berkeley & WEHI)

WW: Wenyi Wang (UT MD Anderson Cancer Center)

AT: Andrew Teschendorff (UCL & CAS-MPG PICB)

**Day 1 (6/24): Statistical methods for analysis of whole-transcriptome data.**

9:30-10:30: Removing unwanted variation (TPS)

10:45-11:45: Removing unwanted variation from RNA-seq data (TPS)

Lunch break

1:30-2:30: Normalizing large, heterogeneous datasets in the omics (TPS)

2:45-3:45: Deconvolution of mixed transcriptomes from tumor samples (WW)

**Day 2 (6/25): Statistical methods for variant calling in high-throughput sequencing data.**

9:30-10:30: Challenges in rare variant calling, what we learned from targeted sequencing studies. (WW)

10:45-11:45: Family-based analysis for calling rare variant in DNA sequencing data (WW)

Lunch break

1:30-2:30: Comparing mutation callers (TPS)

2:45-3:45: Combining mutation callers (TPS)

**Day 3 (6/26). Statistical methods for personalized risk assessment in inherited cancer syndromes.**

9:30-10:30: Mendelian models. (WW)

10:45-11:45: Challenges in application of Mendelian models to various cancer types. (WW)

Lunch break

12:30-3:00 **Lab workshop** on two software packages: BayesMendel, and FamSeq (WW)

**Day 4 (6/27). Statistical Epigenomics**

9:30-10:30: Intra-and inter-sample normalization of Illumina 450k data. (AT)

10:45-11:45: Correction for cellular heterogeneity using devonvolution tools

Lunch break. (AT)

1:30-2:30: General feature selection strategies for EWAS and Cancer Epigenome studies. (AT)

**Day 5 (6/28). Statistical Epigenomics**

9:30-10:30: Dimensional reduction and clustering of DNA methylation data. (AT)

10:45-11:45: Integration of DNA methylation and gene expression data. (AT)

End of the workshop

Software and links:

1. Terry Speed: <http://www.wehi.edu.au/people/terry-speed>
2. Andrew Teschendorff: <https://www.ucl.ac.uk/cancer/rescancerbiol/statisticalgenomics>

For the summer workshop: http://www.picb.ac.cn/compsysg/

1. Wenyi Wang: <http://odin.mdacc.tmc.edu/~wwang7>/

Software:

FamSeq: <http://bioinformatics.mdanderson.org/main/FamSeq>

BayesMendel: <http://bcb.dfci.harvard.edu/bayesmendel>/

**Terry Speed ** A statistician whose primary interest lies in using his expertise to help address problems in genetics and molecular biology. For the past 15 years he has been active in the development of methods for the analysis of microarray data, initially for gene expression, and later for genotyping, exon, tiling and resequencing arrays. Several of the methods he has developed in these contexts have become widely used. With the transition from microarrays to next generation DNA sequencing, his focus has shifted to the development of methods for use with the latter technology. He is currently involved in a number of collaborative projects which seek to throw light on transcriptional regulation by combining the results of ChIP-seq (and before that ChIP-chip) studies involving transcription factors, RNA Pol II and histone marks, with microarray gene expression (and later RNA-seq) data on the same cells. A concurrent interest is in examining the genomic aberrations in tumors using both microarray and next generation sequence data, including mutations, copy number, mRNA and miRNA abundance, and methylation data. In all of the foregoing, he has a particular interest in the broad theme of removal of unwanted variation, due to batch, lab, platform, instrument, operator, environmental conditions and so on.

**Titles and abstracts for five lectures by Terry Speed for Shanghai, June 2015.**

**1. Removing unwanted variation** (jointly with Johann Gagnon-Bartsch and Laurent Jacob.)

Ordinary least-squares is a venerable tool for the analysis of scientific data originating in the work of A-M. Legendre and C. F. Gauss around 1800. Gauss used the method extensively in astronomy and geodesy. In 1907, motivated by social science, G. U. Yule presented a new notation and derived some identities for linear regression and correlation. Generalized least squares is more recent, originating with A. C. Aitken in 1934, though weighted least squares was widely used long before that. At around the same time (1933) H. Hotelling introduced principal components analysis to psychology. Its modern form is the singular value decomposition, (re-)introduced by Eckhart and Young (1936). Random effects models date back to astronomical work in the mid-19th century, but it was through the work of C. R. Henderson and others in animal science in the 1950s that their connexion with generalized least squares was firmly made.

These are the diverse origins of our story, which concerns the removal of unwanted variation in high-dimensional genomic and other “omic” data using negative controls. We start with a linear model that Gauss would recognize, with ordinary least squares in mind, but we add unobserved terms to deal with unwanted variation. A singular value decomposition, one of Yule’s identities, and negative control measurements (here genes) permit the identification of our model. In a surprising twist, our initial solution turns out to be equivalent to a form of generalized least squares. This is the starting point for much of our recent work. In this talk I will try to explain how a rather eclectic mix of familiar statistical ideas can combine with equally familiar notions from biology (negative and positive controls) to give a useful new set of tools for omic data analysis. Other statisticians have come close to the same endpoint from a different perspectives, including Bayesian, sparse linear and random effects models.

**2.** **Removing unwanted variation from RNA-seq data** (jointly with Davide Risso, Lucia Peixoto, and others)

The sequencing of the full transcriptome (RNA-seq) has become the preferred choice for the measurement of genome-wide gene expression. Despite its widespread use, current RNA-seq normalization methods only account for differences in sequencing depth and thus unwanted variation can dominate the signal of interest, leading to low detection power and reproducibility. This is particularly problematic for highly heterogeneous brain tissue. Using RNA-seq data obtained from the mouse hippocampus following learning, we show that commonly used normalization methods are unable to remove artifacts. We present a novel application of the RNA-seq normalization method RUVseq and show that it is able to remove unwanted variation leading to a more accurate and reproducible picture of gene expression changes across different laboratories and behavioral paradigms.

3. **Normalizing large, heterogeneous datasets in the omics** (jointly with Johann Gagnon-Bartsch and Laurent Jacob and Saskia Freytag)

In the omics, large datasets are usually heterogeneous, typically being obtained by aggregating many smaller datasets. For example, the data studied by Kang *et al* (Spatio-temporal transcriptome of the human brain, *Nature* 2011) came from 1,340 samples of post-mortem brain tissue from 57 developing and adult brains taken from 16 regions of the brain at 15 time periods. Each sample gives one microarray gene expression dataset. Large-scale tumour studies, of which there are now many, can exhibit similar heterogeneity. In one such, we have data from nearly 7,000 gene expression microarrays from biopsies on as many breast tumours. For most analyses of such large datasets, some normalization or adjustment will be necessary, where these terms loosely mean dealing with the heterogeneity, whether that is due to technical artefacts in the measurement process, or features of the samples not of interest in the current analysis. The nature of the heterogeneity may be known in part, or quite unknown, and different normalizations will be used for completing different tasks (e.g. comparing, classifying, clustering or correlating). In my talk, I’ll discuss a class of methods we have recently developed for dealing with heterogeneity without assuming the causes are known, methods which make use of what we term negative controls.

**4. Comparing mutation callers (jointly with Su Yeon Kim)**

Somatic mutation-calling based on DNA from matched tumor-normal patient samples is one of thekey tasks carried by many cancer genome projects. One such large-scale project is The Cancer Genome Atlas (TCGA),which is now routinely compiling catalogs of somatic mutations from hundreds of paired tumor-normal DNAexome-sequence datasets. Several mutation-callers are publicly available and more are likely to appear. Nonetheless, mutation-calling is still challenging and there is unlikely to be one established caller that systematically outperforms all others. Evaluation of themutation callers or understanding the sources of discrepancies is not straightforward, since for most tumor studies,validation data based on independent whole-exome DNA sequencing is not available, only partial validation data for aselected (ascertained) subset of sites.

We have analyzed several sets of mutation calling data from TCGA benchmark studies and their partial validation data. To assess the performances of multiple callers, we introduce approaches utilizing the external sequence data to varying degrees, ranging from having independent DNA-seq pairs, RNA-seq for tumor samples only, the original exome-seq pairs only, or none of those.

Utilizing multiple callers can be a powerful way to construct a list of final calls for one’s research. In this first talk, we suppose the availability of RNA-seq data on the same samples, but not independent DNA-seq data.

**5. Combining mutation callers (jointly with Su Yeon Kim).**

Continuing the theme of the previous talk, we now discuss how to use a set of mutations from multiple callers that are impartially validated, i.e. sequenced again independently, under conditions we shall describe. We then present a statistical approach for building a combined caller, which can be applied to combine calls in a wider dataset generated using a similar protocol. The approach allows us to build a combined caller across the full range of stringency levels, which outperforms all of the individual callers.

**Wenyi Wang** Her lab is centered around large-scale complex data sets in recent genomic and familial studies and around important biological questions that emerge from the analysis of these data. Her current research interests include development of methods and software for the accurate measurement of high-throughput genomic data, and development and validation of statistical approaches and software for personalized cancer risk, as well as Bayesian modeling. In particular, she is leading an international collaboration on personalized risk assessment for families with Li-Fraumeni Syndrome, a pediatric syndrome. She also leads a NIH grant on computational deconvolution of gene expression data from heterogeneous tumor samples. Her group recently developed a mutation caller, MuSE, for analyzing DNA sequencing data from matched pairs of tumor-normal samples, currently adopted as one of the main callers for the TCGA sequencing data from 30+ cancer types.

 **Titles and abstracts for five lectures by Wenyi Wang for Shanghai, June 2015.**

**1. Statistical methods for the deconvolution of mixed transcriptomes in heterogeneous tumor samples (1 lecture)** Clinically derived tumor tissues are often times made of both cancer and surrounding stromal cells. The expression measures of these samples are therefore partially derived from the non-tumor cells. This may explain why some previous studies have identified only a fraction of differentially expressed genes between tumor and surrounding tissue samples. What makes the in silico estimation of mixture components difficult is that the percentage of stromal cells varies from one tissue sample to another. Until recently, there has been limited work on statistical methods development that accounts for tumor heterogeneity in gene expression data. I will talk about analytical challenges in this problem and recent progresses made by various groups including ours to address some of them.

**2.** **Statistical methods and tools for variant calling in high-throughput sequencing data** (jointly with Gang Peng, Yu Fan, 2 lectures, 1 lab)

Failure to identify functional variants due to false negatives in variant calling is a challenging problem for many ongoing sequencing studies. Sequencing genomic data from families affected with diseases, such as hereditary cancer syndromes, provides a more tractable approach for finding new functional variants compared to sequencing data from random individuals. However, even in these studies, it is difficult to call variants with a minor allele frequency below 0.1%, as these signals are confounded by technical errors occurring at similar frequencies and are not observed in most other individuals. Considering the still-high cost of sequencing hundreds of disease families, important benefits can be gained from new analytic approaches that guide family-based sequencing studies. I will talk in details about various high-throughput sequencing technology and analysis approaches that have been developed to address the challenges in variant calling.

**3. Personalized risk assessment in cancer** (jointly with Gang Peng, 2 lectures, one lab)

Cancer results from accumulation of multiple genetic mutations. Germline mutation of a cancer gene predisposes the carrier to the development of cancer, known as "inherited susceptibility". This inheritance results in familial clustering of cancers, known as "familial cancer syndromes". Clinical researchers utilize model-based prediction algorithms to identify cancer patients at earlier and more treatable stages and/or to identify healthy individuals at high risk of developing cancer in future. As a result, Mendelian carrier probability models are based on Bayesian methods using detailed family history as input, and have shown performances better than empirical models using regression or classification trees alone. I will talk in details about the development of Mendelian models and their successful application to various cancer types.

** Andrew E. Teschendorff** Principal Investigator at the CAS-MPG Partner Institute for Computational Biology in Shanghai, leading a group in Computational Systems Genomics. He is also an Honorary Research Fellow of the UCL Cancer Institute, University College London. His broad research interest is in Statistical Bioinformatics with an emphasis on Statistical Cancer Epigenomics and Cancer System-omics. Specifically, he is interested in the development and application of statistical methods, notably network physics and signal processing algorithms, to enable a more meaningful interpretation of multidimensional cancer genomic and epigenomic data. The goal is to use computational approaches for hypothesis generation and to help pave the way for improved diagnostic and early detection tools as well as clinically more relevant molecular classifications of common cancers.

**Lectures on Statistical Epigenomics by Andrew E. Teschendorff for Shanghai, June 2015.**

Epigenomics has emerged as a key discipline in our quest to understand molecular biology and the basis of complex diseases like cancer. In our attempt to translate the massive amount of epigenomic data into tangible medical advances, we are faced with many statistical challenges, including normalization, confounding factors, feature selection, intra-sample heterogeneity and integration of multi-dimensional data. The lectures will provide a general introduction to some of the statistical methods being used in Epigenome-Wide Association Studies (EWAS), including Cancer Epigenome studies (TCGA/ICGC). Emphasis will be on DNA methylation as generated with the Illumina 450k platform, although most tools are applicable to other technologies and data types. Topics covered will include (i) Intra-and inter-sample normalization of Illumina 450k data using software tools such as ChAMP, SWAN/BMIQ, COMBAT/SVA/ISVA, (ii) Correction for cellular heterogeneity using devonvolution tools such as RefBasedEWAS, Constrained Projection and RefFreeEWAS, (iii) General feature selection strategies for EWAS and Cancer Epigenome studies, (iv) Dimensional reduction and clustering of DNA methylation data, (v) Integration of DNA methylation and gene expression data, including systems-level approaches such as the FEM algorithm. Throughout the lectures, algorithm usage will be illustrated with examples drawn from large EWAS and Cancer Epigenome studies.