Authors: Gina L. Eagle, Jianguo Zhuang, Rosalind E. Jenkins, John Herbert, Melanie Oates, Francesco Falciani, Neil R. Kitteringham, B Kevin Park, Anna Schuh[1], Stephen Devereux[2], Peter Hillmen[3] and Andrew R. Pettitt

Word count <700 (inclusive of title and abstract text)

Linking genotype to clinical phenotype through proteomic analysis of CLL trial samples

Introduction

Past attempts at understanding the biological basis of chronic lymphocytic leukaemia (CLL) have mainly focussed on genomic alterations and gene expression at the mRNA level. However, the molecular basis of CLL variability remains incompletely understood. We speculate that this is because the clinical phenotype is ultimately determined by gene expression at the protein level. However, to-date there are no large-scale studies which have investigated the human CLL proteome. We have therefore embarked on a large-scale proteomic study of CLL trial samples (~350) for which whole genome sequencing (WGS) data will be available through the Genomics England Ltd (GEL) CLL Pilot Project in an attempt to bridging the gap in our understanding between genotype and clinical phenotype.

Recently developed mass spectrometric (MS) technologies provide an opportunity to develop new diagnostic, prognostic and predictive biomarkers. SWATH (Sequential Windowed Acquisition of all THeoretical fragments) generates a mass spectral library of fragment ions from all detectable peptide precursors. The composite MS/MS spectra are deconvoluted by alignment with a high quality and comprehensive tissue specific database, whereupon patient samples can be stratified based on the quantitative expression profile of thousands of proteins.

We have already generated a CLL-specific database containing mass spectral information for over 7500 proteins found in blood CLL cells. Here we report the biological and technical variability of SWATH data acquired from patient samples and the development of bioinformatic tools to remove batch effects and determine the number of sample preparations and technical replicates required for each patient sample to achieve high statistical power for a large-scale proteogenomic study.

Methods

Protein lysates from 20 cryopreserved CLL PBMC samples (10 IGHV-unmutated and 10 IGHV-mutated) were prepared and delivered into a TripleTOF 6600 mass spectrometer (SCIEX) via an Eksigent nanoLC 415 system (SCIEX). Data independent acquisition was performed using 100 SWATH windows of 5 Da effective isolation width to cover a mass range of 350-1250 m/z. Spectra were aligned using PeakView (SCIEX) against our in-house CLL-specific database (7773 protein entries). Peak areas from peptides with >99% confidence and <1% global false discovery rate were used for protein identification and quantification. A quadratic linear model was used to predict standard deviations at different mean protein levels. This was utilized in a statistical power analysis, to calculate the number of patients needed to reach high statistical power, with a 2 fold change in mean protein expression levels.

To assess the SWATH data for sample preparation and MS run batch effects, 6 cryopreserved CLL PBMC samples (3 IGHV-unmutated and 3 IGHV-mutated) were prepared for SWATH-MS on three separate days, with all prepared samples run on three independent MS runs. This produced a dataset of 54 samples. Variability in and between possible batches were assessed using cluster, PCA, coefficient of variation and ANOVA analyses. Several batch correction algorithms were evaluated for their ability to remove batches.

Results

We have generated a CLL-specific protein database containing 7773 proteins. This covers over 50% of all human UniProtKB/SwissProt entries which have evidence at the protein level (n=14,709), and includes over 87% of the proteins involved in BCR signalling (MetaCore database, Thomson Reuters). Our SWATH-MS data show that biological variation between CLL samples is high, and therefore analysis should be conducted using a large number of trial samples in order to detect differentially expressed proteins with high statistical power. However, technical reproducibility between sample preparations and mass spectrometry runs was shown to be high, with coefficient of variation (CV) of 3.26% and 2.43% respectively. In addition, batch effects were successfully corrected using Combat, from the sva package, in R.

Conclusions

Our results show that number of biological replicates should be prioritised over technical replicates when screening protein expression by SWATH-MS in CLL samples. We are now using these findings and batch correction methods to screen the proteomes of trial samples by SWATH-MS and, in combination with advanced computational biology techniques, will relate protein expression to whole genome sequencing data. In doing so, we hope to identify novel drug targets and predictive biomarkers of disease progression and treatment response.

Affiliations

1. Department of Oncology, University of Oxford, Oxford, UK
2. Department of Haematological Medicine, Kings College Hospital NHS Foundation Trust, London, United Kingdom
3. Faculty of Medicine and Health, University of Leeds, Leeds, UK