# Title Page

# Gene expression analysis in peripheral blood of first episode psychosis patients

Short Title: Expression analysis in blood of psychosis patients

Daniel J Leirer1,2 , Conrad O Iyegbe1,3 , Marta Di Forti1,3 , Hamel Patel1,2 , Elena Carra3, Sara Fraietta3 , Marco Colizzi3, Valeria Mondelli4, Diego Quattrone1 , John Lally3, Olesya Ajnakina3, Sang Hyuck Lee1,2 , Charles J Curtis1,2 , Gerome Breen1,2 , Carmine Pariente4, Katherine Aitchison1, Paola Dazzan1,5 , Robin M Murray3, Richard J B Dobson1,2 and Stephen J Newhouse1,2 \*

Author details:

1 MRC Social, Genetic and Developmental Psychiatry (SGDP) Centre, Institute of Psychiatry, Psychology and Neuroscience, Box P080, De Crespigny Park, SE5 8AF London, UK. 2 NIHR Biomedical Research Centre for Mental Health and Biomedical Research Unit for Dementia and South London at Maudsley NHS Foundation Trust, Box P092, De Crespigny Park, SE5 8AF London, UK. 3 Department of Psychosis Studies, Institute of Psychiatry, Kings College London, Box P092, De Crespigny Park, SE5 8AF London, UK. 4 Department of Psychological Medicine, Institute of Psychiatry, Kings College London, Box P092, De Crespigny Park, SE5 8AF London, UK. 5 Departments of Psychiatry and Medical Genetics, University of Alberta, Alberta, Canada, 11361-87 Avenue, AB T6G 2E1, Edmonton, Canada.

\* Correspondence: stephen.newhouse@kcl.ac.uk 1 MRC Social, Genetic and Developmental Psychiatry (SGDP) Centre, Institute of Psychiatry, Psychology and Neuroscience, Box P080, De Crespigny Park, SE5 8AF London, UK

Transcriptomics; WGCNA; Psychosis; Schizophrenia; Blood; Neurogranin

# Abstract:

**Keywords:**

**Abbreviations:**

# 1.0 Introduction

In recent years Genome Wide Association Studies (GWAS) have resulted in a substantial advance of our understanding of the genetic components to Psychotic Disorders, such as Schizophrenia and Bipolar Disorders (Ripke et al., 2014). Much less focus, however, has been given to high-throughput gene expression analyses in the context of these disorders.

While complementary to GWAS, gene expression microarray studies have the advantage of not just analysing largely static genetic factors, but potentially reflecting dynamic responses to additional factors such as drug use, stress, age and other environmental factors. This is important since we know that psychotic disorders are the result of a complex gene-environment interplay.

An important factor to consider when performing gene expression studies, is the identification of a disorder relevant tissue. For pragmatic reasons, in this study we chose to study transcriptional changes in whole blood, which is easily accessible and minimally invasive. There is an established literature of using blood for gene expression studies of a variety of psychiatric conditions. This includes studies looking specifically at psychosis and / or schizophrenia, however sample sizes in this area have been small, ranging from dozens to about 100 patients (de Jong et al., 2012; Gardiner et al., 2013; Kumarasinghe, Tooney, & Schall, 2012; Kuzman, Medved, Terzic, & Krainc, 2009; Lee et al., 2012; Wu et al., 2016). In addition few studies in this area are directly comparable, due to differences in micro-array platform, and processing of results.

In this study we aimed to identify genes whose transcriptional levels were altered between first episode psychosis patients and controls. We performed a differential gene expression (DGE) analysis, followed by gene enrichment analysis and network analysis.

# 2.0 Methods and Materials

## 2.1 Ethics

The Study received ethical approval from the South London and Maudsley (SLaM), as well as from the Institute of Psychiatry Local Research Ethics Committee, IOP/SLAM research ethics approval number: 135/05. Informed written consent was obtained from all participants in the study prior to start.

## 2.2 Study design and participants

As part of the GAP study (Di Forti et al., 2009, 2015). We approached all patients aged 18–65 years who presented with first-episode psychosis at the inpatient units of SLaM were approached. We invited patients to participate if they met the International Classification of Diseases 10 criteria for a diagnosis of non-affective (F20–F29) or affective (F30–F33) psychosis, validated by administration of the Schedules for Clinical Assessment in Neuropsychiatry (SCAN). We excluded individuals who met the criteria for organic psychosis (F09). If patients were too unwell to cooperate, we re-contacted them after the start of treatment. Between May 1, 2005, and May 31, 2011, we recruited 461 patients with first-episode psychosis. The cohort consisted of a diverse multi-ethnic population. Further patient information, blood samples and genetic ancestry where acquired as described previously (Di Forti et al., 2012). During the same period, we recruited 389 control individuals, aged 18–65 years, who were similar to the local population in terms of gender, ethnic origin, education, and employment status, and socio-economic status. We recruited controls using internet and newspaper advertisements and by distributing leaflets at train stations, shops, and job centres. Volunteers were administered the Psychosis Screening Questionnaire (Bebbington & Nayani, 1996) and were excluded if they met the criteria for a psychotic disorder or if they reported a previous diagnosis of psychotic illness.

## 2.3 RNA processing and Quality Control

Whole blood samples were collected using PAXgene tubes for RNA, from a subset of GAP participants (227 cases and 168 controls). Psychotic patients were stabilized using anti-psychotics for a week. Samples were run at the NIHR Biomedical Research Centre for Mental Healthy (BRC-MH) microarray facility at the SGDP, Institute of Psychiatry, and King’s College London. Microarrays where run in accordance with the manufacturer’s protocol using Illumina HT-12 V4 beadchips (Illumina, USA).

All analysis was performed using R version 3.1.2 (Team, 2013). We performed rigorous quality control, by pre-processing the data using an adapted in-house developed pipeline (https://github.com/snewhouse/BRC\_MH\_Bioinformatics).

The pipeline takes raw gene expression data exported from Illumina’s Genomestudio, performs background correction (Xie, Wang, & Story, 2009) using negative bead expression levels in order to correct for noise. Lumi (version 2.22.1 (Du, Kibbe, & Lin, 2008)) was used to log base 2 transform the data followed by robust spline normalization (Du et al., 2008). Outlying samples were iteratively identified using fundamental network concepts and removed, following the methods described by Oldham et al. (Oldham, Langfelder, & Horvath, 2012).

In order to reduce the influence of batch effects we identified significant confounding variables by using the first principle component of housekeeping and undetected probes and regressing this against technical variables. In cases where the variables were significantly associated with the first principle component, they were regressed against expression for each probe, and the mean adjusted residuals were taken forward. The resulting adjusted expression matrix was subjected to surrogate variable analysis, using the SVA package (Leek, Johnson, Parker, Jaffe, & Storey, 2012), to identify potential unknown batch effects. Following this we compared recorded gender with gender determined by XIST and PRKY probes, and excluded samples that showed a mismatch. Finally, we excluded all probes that could not be reliably detected in 80% of the samples in at least one diagnostic group. We used the R package CellMix version 1.6 (Gaujoux & Seoighe, 2013), to test for potential significant differences in whole blood cell populations between cases and controls. Prior to further analysis we controlled for CellMix derived cell proportions, Age, Gender and Ethnicity using a linear model to create an adjusted expression matrix.

## 2.4 Differential gene expression analysis

To identify differentially expressed genes (DE), the R package LIMMA (Smyth, 2004) (version

3.26.8) was used. Cell proportions, age, sex and ethnicity were previously regressed out. Probes were declared significantly differentially expressed if the FDR (false discovery rate) adjusted q-value was less than 0.05 and the absolute log fold change was above 0.1. Probes annotated with the “LOC” or ”HS.” prefix were filtered out at this stage.

## 2.5 Weighted Gene Co-expression Network Analysis

In order to identify modules based on co-expression we used the WGCNA R package (Langfelder & Horvath, 2008). For this analysis we filtered out duplicate probes mapping to the same gene. An adjacency matrix was generated using a β of 6 which met the scale-free topology criteria. A hierarchical clustering tree was created and modules were originally defined using the WGCNA function cutreeDynamic with a minimum module size of 20. Modules were then merged using the mergeCloseModules function with a threshold value of 0.25. The eigengene of each module was then correlated with phenotypic information.

## 2.6 Gene enrichment analysis

All enrichment analysis was performed using the UserListEnrichment function in R (https://cran.r-project.org/web/packages/WGCNA/index.html). This function is part of the WGCNA package (Langfelder & Horvath, 2008). Enrichment analysis, for the results of the LIMMA analysis, was performed by testing differentially expressed probes (q-value =< 0.05 and ± logFC > 0.1). All probes that did not pass the q-value threshold were included as background.

We also submitted probe lists corresponding to the modules we found in our WGCNA analysis for enrichment analysis. This was done by using core genes of each module which were defined as probes with an above average module membership. All other probes were labelled as background.

In all cases we used KEGG (M. Kanehisa & Goto, 2000; Minoru Kanehisa, Goto, Kawashima, Okuno, & Hattori, 2004) and GO (Ashburner et al., 2000) databases. The databases we used were ”KEGG 2016”, ”GO Molecular Function 2015”, ”GO Cellular Component” and ”GO Biological Process 2015”. We downloaded these databases from the Enrichr website (http://amp.pharm.mssm.edu/Enrichr/#stats, accessed 05.July.2016). Result categories that contained less than 5 overlapping probes were filtered out.

## 2.7 Post hoc analysis of medication

A post hoc analysis using antipsychotic data was incorporated to the differential expression and WGCNA analysis, in order to address if observed changes could be entirely attributed to medication. We split the patient data into 4 sub-groups corresponding to medication status at the time of blood collection. The groups were Antipsychotic Naïve, only Risperidone, only Olanzapine and one group that included all patients with recorded Antipsychotic prescription. These groups were compared with controls and incorporated to our Limma analysis by using previously identified and significantly differentially expressed probes. In our WGCNA we correlated the module eigengenes with these comparison groups.

# 3.0 Results

## 3.1 Sample Characteristics

Out of the 395 original samples (227 cases and 168 controls), 280 samples passed quality control, and had full information on Age, Gender and Ethnicity. This corresponded to a final population of 131 first episode psychosis cases and 149 controls. The basic demographics of the final sample population are shown in Table 1. Patients were less likely to be Caucasian (p-value = 0.049) and more likely to be smokers (p-value =< 0.001). There was no significant difference in age, gender or body mass index.

Antipsychotic medication status at the time the blood was taken, and icd10 diagnosis are show in Table 2. 18 Patients (13.7%) were unmedicated when blood samples were taken, and 13 patients (9.9%) had unavailable information on medication for this study. The remaining patients, were primarily medicated with Olanzapine (35.1%), or Risperidone (20.6%). The next most common medications were Aripiprazole (9.2%), Haloperidol (4.6%) and Quetiapine (4.6%). 3 other drugs were taken by a single individual.

In terms of icd10 diagnosis, 50.4% of patients had an ultimate diagnosis of schizophrenia. 12.2% had a diagnosis of mania with psychosis, 6.9% had a diagnosis schizoaffective disorder, 7.6% had various forms of depression, 3.8% had delusional disorder, 1.5% bipolar disorder and 17.5% had incomplete records, did not meet any criteria, or had unspecified psychosis.

## 3.2 Differential Expression analysis

Out of the total 4730 probes remaining after quality control, 667 were removed due to being considered poorly annotated (LOC or HS. prefix), leaving a total of 4062 probes (see Supplementary Table 1, for full probelist with columns indicating, removed probes, significant probes, and background probes). In this reduced probeset we found 877 significantly DE genes (q-values < 0.05 & absolute logFC > 0.1). Out of these, 460 were up-regulated, and 417 probes were down-regulated (Supplementary Table 2). The top 50 up and downregulated probes (q-values < 0.05) in terms of fold change are shown in Table 3 and 4. The most significant up and downregulate probes were, SUMO3 (logFC =0.1, q-value = 7.36E-05) and HNRNPUL2 (logFC =-0.18, q-value = 0.000206) respectively. The significant probes with the highest and lowest logFC were DEFA1B (logFC =0.91, q-value = 0.00021) and VWCE (logFC =-0.37, q-value = 0.00394) respectively.

## 3.3 Weighted Gene Co-expression Network Analysis

Weighted Gene Co-expression Network Analysis identified 14 modules (See Supplementary Table 3 for module membership of each probe). The network dendrogram is show in Supplementary Figure 1. We tested the module eigengenes for their correlation with traits including case control status, BMI and Tobacco use. We also tested the correlation between control status and Antipsychotic Free patients, as well as patients on either Risperidone or Olanzapine.

Seven were found to be statistically significant for case-control status. These were the Turquoise, Brown, Blue, Green, Black, Yellow and Magenta modules (see Supplementary Figure 2, p-value < 0.05). The most significant of these modules were the Turquoise (R2 =-0.30, p-value= 5x10−8, Size = 1484 probes), Brown (R2 =-0.22, p-value= 9x10−5) and Blue (R2 =0.2, p-value= 4x10−4) modules. These modules respectively contained 1484, 424, 837 probes (Supplementary Table 2 for probe list, Supplementary Figure 1 for relative sizes).

## 3.4 Functional Enrichment analysis

Copy and edit heavily

## 3.5 Post hoc analysis of medication

Describe results. Namely in reduced data, we see large statistically significant differences between AF and Control, and Med vs Control. Med vs Control, is more, AF has more logFC spread.

# 4.0 Discussion

## 4.1 Glutamate

Rewrite and purge

## 4.2 Platelets and DiGeorge syndrome

Rewrite and purge

## 4.3 Defensins

Rewrite and purge

## 4.4 WGCNA

Rewrite and purge

## 4.5 Enrichment Results

Rewrite and purge

## 4.6 Confounding by medication

Summary of WGCNA and volcanoplots. Note medication results are limited and cannot fully be addressed. Spread is more pronunced in AF than antipsychotics. May indicate effect of antipsychotics. Too noisy to be able to eliminate the possibility that medication has no effect. But the authors interpretation is that an effect is liekly which is backed by the literature., The extent of this is debatable.

Mention directionality in modules. BMI and Tobacco are included.

## 4.7 Conclusion and Limitations

Copy and edit

# Acknowledgements:

# Financial Disclosures:

# References:

# Table/Figure Legends: