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

Short Title: Expression analysis in blood of psychosis patients

**Background:**

Psychosis is a complex phenotype representing the interplay between genes and environment. It is associated with a number of psychiatric disorders, most notably schizophrenia and bipolar disorder. The disorder manifests as any combination of hallucinations, delusions, catatonia, or thought disorder, and can have a severe detrimental impact on quality of life and life expectancy. In this study we performed whole-genome transcriptional profiling to identify genes and pathways associated with psychosis.

**Methods:**

A total of 150 controls and 163 first episode psychosis cases were recruited and had their peripheral blood gene expression profiles measured using the Illumina Human HT\-12\-v4 microarray platform. We identified differentially expressed genes (adjusting for age, sex and ethnicity), and report networks of co-expressed genes and pathways associated with psychosis.

**Results:**

Among our findings we report an enrichment for the Glutamate system and Post-synaptic density genes. Notably we find a number of biologically interesting genes such as Neurogranin (NRGN) and Septin 5 (SEPT5) in these pathways which have been previously linked to psychosis. We hypothesis that these genes play an important role in platelet function as well as in neurotransmitter signalling, and that disruptions in these pathways may therefore be detected in several tissues.

**Conclusions:**

Our results are consistent with the Glutamate hypothesis, and suggest that subtle signals, indicative of the underlying biological changes in the central nervous system of psychosis patients, may be characterised in peripheral tissue.

Transcriptomics; WGCNA; Psychosis; Schizophrenia; Blood; Neurogranin

# 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 [1]. 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 [2–7]. 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.

# Methods and Materials

## 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.

## Study design and participants

As part of the GAP study [8, 9]. 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 [10]. 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 [11] and were excluded if they met the criteria for a psychotic disorder or if they reported a previous diagnosis of psychotic illness.

## 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 [12]. We performed rigorous quality control, by pre-processing the data using an 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 [13] using negative bead expression levels in order to correct for noise. Lumi (version 2.22.1 [14]) was used to log base 2 transform the data followed by robust spline normalization [14]. Outlying samples were iteratively identified using fundamental network concepts and removed, following the methods described by Oldham et al. [15].

In order to reduce the influence of batch effects we adjusted for technical categorical variables using a combination of robust linear regression [16] and ComBat [17]. Other Continuous technical artefacts were accounted for by taking the first principal component across housekeeping and undetected probes and regressing this against technical variables. In cases where the variables where significantly associated with PC1 they were regressed against expression for each probe, and the mean adjusted residuals where taken forward for all further analysis. 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 [18], to test for potential significant differences in whole blood cell populations between cases and controls. None were found.

## Differential gene expression analysis

To identify differentially expressed genes (DE), the R package LIMMA [19] (version

3.26.8) was used. Age, sex and ethnicity were included as covariates. Probes were declared significantly differentially expressed if the FDR (false discovery rate) adjusted q-value was less than 0.05. Highly significantly differentially expressed probes were defined as probes with a >10% fold change and an FDR q-value less than 0.05. Probes annotated with the “LOC” or ”HS.” prefix were filtered out at this stage.

## Weighted Gene Co-expression Network Analysis

In order to identify modules based on co-expression we used the WGCNA R package [20]. 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 30. 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.

## 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 [20]. Enrichment analysis, for the results of the LIMMA analysis, was performed by testing differentially expressed probes (q-value <= 0.05). 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 (defined below), and all other probes as background. Core genes for modules strongly associated with case control status (p-value =< 0.001) were defined as probes with an above average module membership (average of MM for all probes in module) and an above average correlation with psychosis. Core genes for all other modules were defined as probes with an above average module membership.

In all cases we used the internal Brain, Blood and Immune Userlistenrichment libraries. In addition we manually added KEGG and GO databases to the function. The databases we used were ”KEGG 2016”, ”GO Molecular Function 2015” 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.

# Results

## Sample Characteristics

Out of the 395 original samples (227 cases and 168 controls), 313 samples passed quality control, and had full information on Age, Gender and Ethnicity. This corresponded to a final population of 163 first episode psychosis cases and 150 controls. The basic demographics of the final sample population are shown in Table 1. Overall the patient group was slightly younger, contained a higher percentage of males and a lower percentage of Caucasians, compared to the control group. These differences were found not to be statistically significant in this sample.

## Differential Expression analysis

Out of the total 4745 probes remaining after quality control, 676 were removed due to being considered poorly annotated (LOC or HS. prefix), leaving a total of 4069 probes (Supplementary Table 1). In this reduced probeset we found 2199 significantly DE genes (q-values < 0.05). Out of these, 828 were up-regulated, and

1371 probes were down-regulated (Supplementary Table 1).

The top 50 up and downregulated probes (q-values < 0.05) in terms of fold change are shown in Table 2 and 3. The top upregulated probe was DEFA1B (logFC =

0.65, q-value = 0.00017) and the top downregulated probe was VWCE (logFC =

-0.44, q-value = 0.00009).

## Weighted Gene Co-expression Network Analysis

Weighted Gene Co-expression Network Analysis identified 11 modules (See Supplementary Table 2 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. 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).

## Functional Enrichment analysis

We performed enrichment analysis on our full list of DE (N=2199) genes, see Table

3 for top 15 results and Supplementary Table 3 for the full list of 198 categories that passed a p-value threshold of 0.05. Out of these 198 categories only 11 passed a multiple testing threshold of 0.05. These categories represent functions involved in Translation, protein targeting and localization to the Endoplasmic Reticulum or Membrane, viral transcription and GlutamatergicSynapse signalling. With the exceptions of the GlutamatergicSynapse result all significant categories are part of the GO biological processes library (see Supplementary Table 3, Type = User, Category

= \*GO BP) and are highly enriched for probes mapping to ribosomal proteins. The most significant result was ”cotranslational protein targeting to membrane” (q-value

= 0.000000046). The ”GlutamatergicSynapse (Mouse)” (q-value = 0.005) result was the 7th most significant results, and contained the NRGN probe. Other categories that failed to reach our q-value threshold, but were among the top results included

”PostSynapticDensity proteins”, “Glutamatergic Synaptic Functions”, and categories associated to Platelets and the Mitochondria. The Glutamatergic and Post synaptic density modules are part of the brain atlas library and are associated to publications ([21, 22]).

We also aimed to identify pathways associated to the WGCNA modules we found. For this purpose we also used the UserListEnrichment, and only used core genes for each module (Supplementary Table 2). Enrichment analysis of the WGCNA modules identified pathways in 9 of the 11 modules (Table 4, Supplementary Table 4). The Grey module was excluded due to being an aggregate of unassigned probes. The Magenta module did not contain any gene enrichment categories with a p-value below 0.05 and a minimum of 5 probes. However the Modules most highly correlated with Psychosis (Turquoise, Brown and Blue) all failed to reach statistical significance for any category when accounting for multiple testing (Supplementary Table 4). The most significant categories for the Turquoise Module were ”Glycolysis / Gluconeogenesis” (p-value = 0.00025, q-value = 1), ”gluconeogenesis” (p-value = 0.00075, q-value = 1). For the Brown module they were ”GlutatmatergicSynapse” (p-value = 0.0024, q-value = 1) and ”GlutamatergicSynapticFunction” (p-value = 0.047, q-value = 1). For the Blue module they were ”RedBloodCell” (p-value = 0.00000082, q-value = 0.052) and ”response to lipopolysaccharide” (p-value = 0.0000037, q-value = 0.24). The Yellow module which showed some association with psychosis (p-value = 0.007) was characterised by Pathways associated to Metabolism, the Cell Cycle and Autism, but also failed to pass the q-value threshold. The 5 remaining modules contained at least one category that was significant after controlling for multiple testing. These are the Black, Green, Greenyellow, Pink, and Purple modules. Out of these the Black and Green modules are subtly correlated to psychosis (p-values = 0.003 and 0.002 respectively, Supplementary Figure 2). The Black module was enriched for Ribosomal proteins and Translation related Pathways, while the Green Module contained a number of Genes associated to the Immune System. The Greenyellow, Pink and Purple modules were associated to Immune Cells, the Cell Cycle and Cytokine signalling respectively.

# Discussion

We report to our knowledge the largest blood based transcriptomic study conducted for first episode psychosis. We found a much larger number of differentially expressed and statistically significant probes compared to previous studies [2, 3]. This may be due, in part to our larger sample size and our stricter pre-processing pipeline, which resulted in fewer probes passing quality control and thus reducing the multiple testing burden.

## Glutamate

Notably we found pathways and probes that are significantly associated with the glutamate system which is at the centre of prevailing hypotheses surrounding schizophrenia and psychosis (Table 3). When looking at the probes represented in the Glutamate and post synaptic density modules (Supplementary Table 4), we found a number of probes, which were also represented in our top differentially expressed probes. Among them were a number of ribosomal proteins which mapped to the Post Synaptic Density Protein category. This may indicate that the ribosomal signal is in part associated to the nervous system.

Notably we also identified NRGN (neurogranin (protein kinase C substrate, RC3) in the Glutamate module. NRGN is especially interesting since it was one of the early successful schizophrenia GWAS hits ([23]). Further it has been reported that NRGN protein levels are lower in the prefrontal cortex of schizophrenia sufferers ([24]), which is consistent with the lower NRGN levels found in our data. NRGN in addition has a biologically plausible mechanism in which it is related to schizophrenia, by being oxidized as a direct response to glutamate stimulation (see Stefanson et al for further details ([23])). Overall NRGN was the 5th most downregulated probe we identified (logFC = -0.37, Q-value = 0.000013).

## Platelets and DiGeorge syndrome

While performing our analysis we were cautious to account for potential variations in cell proportions between cases and controls. Our main approach for this was to use CellMix, but we also included a blood specific library (Supplementary Table 3, Type = Blood) in our enrichment analysis. The only significant association we found in this library, was ”BloodPlatelets Top50 Gnatenko” (p-value =

0.0068). While this result does not stand up to multiple testing (q-value = 1), it is interesting to note that the platelet module contained NRGN, which we also found in the Glutamate pathways. Indeed the original study that this platelet category is based on ([25]) found NRGN to be the second most highly expressed gene in platelets. In line with this the platelet specific Glycoprotein IX (GP9) was the 3rd most downregulated probe in our data (logFC = -0.39, q-value = 0.000012). GP9, together with GPlb, forms the GPIb-IX-V complex which acts exclusively on the surface of platelets. This might suggest an association of platelet dysfunction in psychosis patients. A genetic susceptibility to low NRGN expression may also affect platelets, considering the seemingly high abundance of NRGN in platelets under normal circumstances. NRGN and GP9 also cluster together in the top left of the volcano plot (Figure 1) with one other probe, namely Septin 5 (SEPT5, logFC =-0.35, q-value = 0.000012). Septin 5 (also known as CDCrel-1) is the 7th most downregulated probe we identified and plays a role in vesicle fusion, via the SNARE (soluble N-ethylmaleimidesensitive fusion protein attachment protein receptor) complex [26]. This is important in processes such as Neurotransmitter release and autophagy, and has recently been implicated in a differential expression study of schizophrenia [6]. Septin 5 is also a parkin substrate implicated in Parkinson’s disease ([27]). Interestingly Septin 5 has been shown to accumulate in parkin knockout mice which displayed motor dysfunction symptoms similar to Parkinson’s disease ([27]). This was attributed to dopaminergic cell death in the mice. This accumulation of Septin 5 is consistent with the prevailing dopamine hypothesis regarding Parkinson’s disease. Since antipsychotic use can often lead to symptoms mimicking Parkinson’s disease, there is some plausibility for this biological connection, especially since we see a down regulation of Septin 5 here.

In addition to this SEPT5 has been shown to be expressed in platelets. One study showed that CDCrel-1 (i.e. SEPT5) Null mouse mutants had platelets that released stored serotonin at lower stimulation thresholds than control animals ([28]). Low SEPT5 levels may predispose cells to release higher levels of Serotonin and Dopamine at lower stimulation. The authors mentioned that the SEPTIN family likely shows a high degree of redundancy, which would reduce the observed impact of a single gene deregulation.

Interestingly, SEPT5 is located in the 22q11.2 region, in the same genomic area as Catechol-O-methyl transferase (COMT) which has been implicated in schizophrenia susceptibility due to its role in dopamine turnover in the frontal cortex ([29]). When this region is deleted it leads to DiGeorge or 22q11.2 deletion syndrome ([30]). People with this syndrome have an especially high rate of psychotic disorders such as schizophrenia. One study found that almost 1 in 3 adults with this deletion suffered some psychotic disorder ([31]). What is notably is that individuals with

22q11.2 deletion syndrome also often have platelet disorders [32] which are likely associated to the deletion of GPIb-β. GPIb is located directly next to SEPT5 in the genome. So close in fact, that they often create read through mRNA transcripts. As mentioned before GPlb binds to GPIb to form the GPIb-IX-V receptor complex, which is important for normal platelet functioning.

Taken together, this might suggest that psychosis patients with a genetic predisposition to disturbances in NRGN or SEPT 5 related signalling pathways in the CNS, may also be more vulnerable to deregulation in similar pathways in periphery tissues such as blood. This correlation between SEPT5, GP9 and NRGN is further corroborated by our WGCNA results which place all of them in the Turquoise module (Supplementary Table 2).

## Defensins

Another interesting result is the up regulation of Alpha Defensin probes in psychosis patients, which represent the 5 highest upregulated probes in psychosis patients in this study. This is consistent with a previous microarray study of similar size looking at schizophrenia patients ([3]) as well as 2 proteomic studies of Schizophrenia (Plasma) and Bipolar (Saliva) patients ([33, 34]). Alpha Defensins are primarily involved in the innate immune system specifically in viral opsonisation. This may have some peripheral link to psychosis, although in the absence of a plausible mechanism linking them to psychiatric conditions they may prove useful as a biomarker that can be incorporated into classification models. In addition to this finding replicates previous findings, which strengths our results overall.

## WGCNA

The WGCNA results identified pathways associated to 9 modules (Table 4, Supplementary Table 4). The modules most strongly associated to psychosis were the Brown, Turquoise and Blue modules.

The Brown module was notably enriched for Glutamate pathways, which were identified in a mouse and human study [21, 35]. These result should be interpreted cautiously, due to each category only containing 7 probes and not passing a multiple testing threshold (p-values = 0.002 and 0.04, q-values = 1 and 1). However, It is notable that both categories contain the NMDAR subunit GRINA (Glutamate Ionotropic Receptor NMDA Type Subunit Associated Protein 1). GRINA is located in chromosome 8q24.3, an area prone to CNVs, were it has been implicated in epilepsy [36], mental retardation [37] and depression [38]. GRINA was downregulated in Psychosis patients in our differential expression analysis (logFC = -0.16, q-value

= 0.007). GRINA is expressed in blood, and due to its location in a CNV prone area its level of expression may give an indication of disruptions in glutamate signalling in the CNS.

The Turquoise module is by far the largest module we found. It contains 1484 probes, including the previously mentioned probes for NRGN, GP9 and SEPT5. It represents a large part of all the probes we identified, which makes it difficult to assign a concrete and specific molecular function to this module. Gene enrichment analysis identified no pathways associated to this module, after correction for multiple testing (Supplementary table 4). However the top categories for the module were related to Gluconeogenesis (Table 4). Notable here was the recurrent GAPDH, PGAM1, PGAM4, PGK1 probe signature. These probes are directly involved in Glycolysis and Gluconeogenesis, which is a mechanism that has been reported several times in the literature in relation to Psychosis [39, 40]. These studies also reported downregulation of 14-3-3 proteins, notably YWHAZ and YWHAH. Both of these proteins are members of the Turquoise module and are subtly downregulated in our data (logFC -0.063 and -0.069). Despite the role of NRGN in glutamate signalling, we found no pathways in this module associated to it. This might be because the majority of the glutamate signal appears to be in the Brown module.

The Blue module is primarily enriched for pathways associated to the immune system. The immune system has long been implicated in Schizophrenia and other Neuropsychiatric disorders [1, 23, 41, and 42]. This has been further corroborated by gene expression studies [3–6]. Of particular interest here are the Chemokine pathways, which were found in a previous gene expression study of similar size [3].

## Enrichment Results

Other Gene Enrichment results, both from WGCNA modules and our DGE analysis, contained pathways associated to the Ribosome, Translation, Protein localization (DGE results and Black Module) and Metabolism (Yellow Module). While these pathways in addition to Glucose Metabolism (Turquoise Module), and Immune Pathways (Blue Module) have previously been associated to Schizophrenia, they have also been associated to a number of other neurological and psychiatric disorders, such as Alzheimer’s disease [43], Parkinson’s disease [44], Depression [45] and Autism [46] among others. Further a 2009 review of 44 proteomic studies identified energy metabolism as well as protein synthesis among the most commonly identified genes and pathways [47]. It may be that perturbations in these systems represent more general patterns associated to stress and ill health, rather than signals specific to psychosis.

## Conclusion and Limitations

Overall we present the results of GAP gene expression study. We report that our differential expression analysis partially replicates and expands on previous findings in the literature. We found evidence for differential expression of probes involved in Neurotransmitter signalling (NRGN, SEPT5, GRINA) and the immune system (Defensins), and identified pathways associated to Glutamate signalling, Metabolism and the Immune System. While these results seem to be consistent with established results and hypotheses in the literature, they should be interpreted cautiously. There are a variety of confounding factors we were not able to address adequately, among these would be smoking and medication, which are likely to have profound effects on gene expression. Despite this we show evidence of plausible gene expression signatures for psychosis in blood, and while these results need to be replicated they are highly encouraging for follow up studies.