Personalized Perturbation Profiles Reveal Concordance between Autism Blood Transcriptome Datasets

Jason Laird[1] & Alexandra Maertens[1,2]

[1]Rabb School of Continuing Studies, Brandeis University, Waltham, MA, USA
[2]Department of Environmental Health and Engineering, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

Abstract

The complex heterogeneity of Autism Spectrum Disorder (ASD) has made quantifying disease specific molecular changes a challenge. Blood based transcriptomic assays have been performed to isolate these molecular changes and provide biomarkers to aid in ASD diagnoses, etiological understanding, and potential treatment [1–6]. However, establishing concordance amongst these studies is made difficult in part by the variation in methods used to call putative biomarkers. Here we use personal perturbation profiles to establish concordance amongst these datasets and reveal a pool of 1,189 commonly perturbed genes and new insights into poorly characterized genes that are perturbed in ASD subjects. We find the resultant perturbed gene pools to include the following unnamed genes: C18orf25, C15orf39, C1orf109, C1orf43, C19orf12, C6orf106, C3orf58, C19orf53, C17orf80, C4orf33, C21orf2, C10orf2, C1orf162, C10orf25 and C10orf90. Investigation into these genes using differential correlation analysis and the text mining tool Chilibot reveal interesting connections to DNA damage, ubiquitination, R-loops, autophagy, and mitochondrial damage. Our results support evidence that these cellular events are relevant to ASD molecular mechanisms. The personalized perturbation profile analysis scheme, as described in this work, offers a promising way to establish concordance between seemingly discordant expression datasets and expose the relevance of new genes in disease.

Introduction

ASD is an incredibly complex disease marked by an even more complex genetic presentation. The Simons Foundation Autism Research Initiative SFARI has curated a list of over 900 genes implicated in ASD [7]. Males seem more likely to develop ASD, with Loomes et al. 2017 noting the ASD male to female ratio is close to 3:1 [8]. This discrepancy between males and females is not currently understood [9]. What is more, there is an extreme bias towards males in autism research with Ratto et al. 2019 noting that female specific ASD traits are being missed by behavioral diagnosis [10]. Multiple transcriptomic studies have been performed with the aim of supplementing psychiatric data with quantification of the molecular differences between typically developing individuals and individuals with ASD [1,4–6,11,12]. The transcriptomic assays examined in this work are blood-based, which offers the advantage of being easily available but the disadvantage of only revealing indirectly the likely phenotypic changes seen in the brain. Brain biopsies are more difficult to obtain, and are typically a complex mix of cells, whereas blood is more readily available and can be tracked over time. By exploring aberrant genetic expression in blood, we aim to offer a more robust understanding of ASD molecular mechanics in a tissue well suited for sampling.

Instead of traditional methods of establishing differential expression, this work uses "personal perturbation profiles" - wherein a z-score based on the expression values of the control group determines whether a gene in an ASD subject is perturbed [13]. Personal perturbation profiles are an admittedly interesting choice for a concordance study as Menche et al. 2017 showed these profiles often have very little overlap, even in the same study[13]. However, traditional

methods of establishing differential expression are not without drawbacks. Isolating a common pool of biomarkers across studies is often difficult due to the variety in platforms and statistics used in delivering these results [13] . Balázsi et al. 2011 point out biological processes are marked by random fluctuations providing an inherent source of noise when analyzing biological data [14] . While these concerns are not explicitly resolved with personal perturbation profiles, they do attempt to account for another source of biomarker variation between studies, disease heterogeneity and a disease that results from a multitude of molecular mechanisms[13] . Given the apparent heterogeneity of ASD, the personal perturbation approach seemed appropriate.

This work aims to reveal not only a set of commonly perturbed genes implicated in ASD but explore poorly characterized genes that are commonly perturbed in ASD individuals. We selected genes with no canonical name at the time these studies were performed. These genes were isolated by the pattern "CXXorfXXX" in the gene name. While isolating genes based on this pattern is a relatively quick and simple way of isolating poorly characterized genes, it has some advantages over other methods - for example, we could have selected genes based on PMID count or GO annotations. However, the majority of GO annotations are held by only 16% of all known genes, [15] and these annotations are used as evidence in many articles which may inflate literature counts. Additionally, not all genes are mapped to PMIDs, nor is there any obvious cut-off as to what makes a gene poorly studied. For the sake of a first pass analysis, we chose to pick unnamed genes to isolate a smaller pool of poorly characterized candidates.

Results

Personalized Perturbation Profiles Reveal a Set of Commonly Perturbed Genes

The following ASD blood-based transcriptomic studies were examined, GSE26415,GSE6575,GSE37772, GSE18123-GPL570,GSE18123-GPL6244,GSE42133, and GSE25507. Data was pulled from the Gene Expression Omnibus GEO. The tissue source used in GSE26415, GSE6575, GSE37772, GSE18123-GPL570,GSE18123-GPL6244, GSE42133, and GSE25507 were venous leukocytes, whole blood, lymphoblast cell line, blood, blood, leukocyte, and lymphocytes, respectively. Previous studies into ASD blood-based transcriptomics have shown via multidimensional scaling that there are notable differences between these tissue types [16] . However, the availability of large blood based ASD datasets has limited us to these datasets in this first attempt to establish concordance.

We first began by taking a more traditional approach to differential expression, by using the Welch two sample t-test to isolate which genes were differentially expressed in each dataset. T-test p-values were limited to those below the canonical 0.05 threshold to reveal each dataset had an average of 4,436 differentially expressed genes. When each dataset's differentially expressed genes were compared with one another, these datasets only had one gene in common, SMARCA2.

To establish personal perturbation profiles, the expression of each gene for each ASD individual was compared to the expression values of the controls for that gene per study. If the z-score of the expression value for the ASD individual was greater than 2.5 or less than -2.5 the gene was considered perturbed. The list of perturbed genes per patient was added to the list of genes perturbed in that study (**Supplemental Data 1**). On average there were 14,117 unique personally perturbed genes per study **Figure 1**. These pools were compared with one another to reveal 1189 genes in common **Figure 1** (**Supplemental Data 2**). In total, there were 534 ASD individuals between the studies examined. Only 9.2% of these ASD individuals were female and the remaining 90.8% were male. The extreme bias toward males in these data prompted us to account for gender. To obtain a pool of male specific personally perturbed genes each study's list of personally perturbed genes was limited to the genes found to be perturbed in male ASD patients. These male only perturbed gene pools were compared with one another to obtain a pool of commonly perturbed male genes. The pool of commonly perturbed female genes was obtained using this same scheme. ASD

males were found to have 1,111 commonly perturbed genes (**Supplemental Data 2**), while ASD females were found to have 133 commonly perturbed genes (**Supplemental Data 2**).

Figure 1
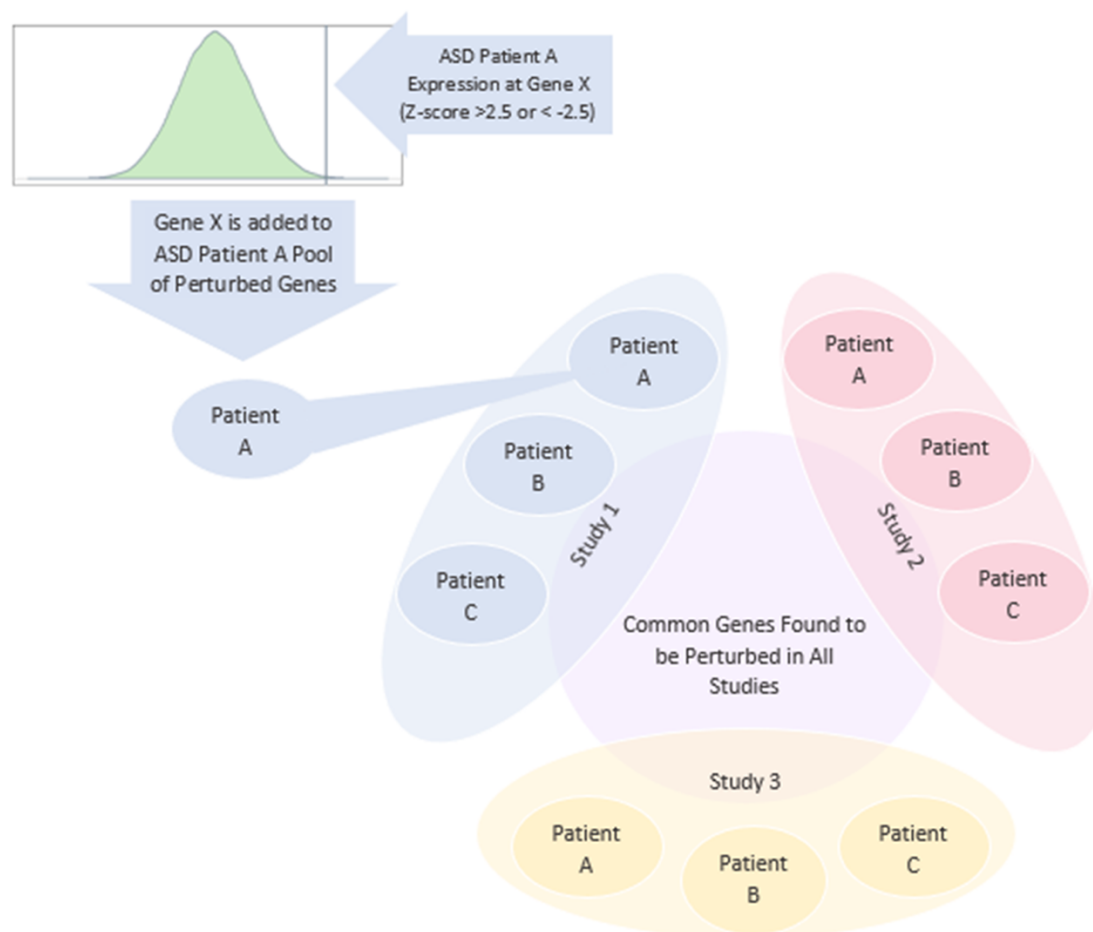
Pools of personally perturbed genes per study reveal a set of commonly perturbed genes. The expression of each gene for each ASD individual was compared to the expression values of the controls for that gene per study. If the z-score for the expression of the ASD individual was greater than 2.5 or less than -2.5 the gene was considered perturbed. The list of perturbed genes per patient was added to the list of genes perturbed in that study. These pools of genes were compared with one another to reveal a set of 1189 commonly perturbed genes, 1,111 genes when each study's pool was limited to males and 133 when the pool was limited to females.

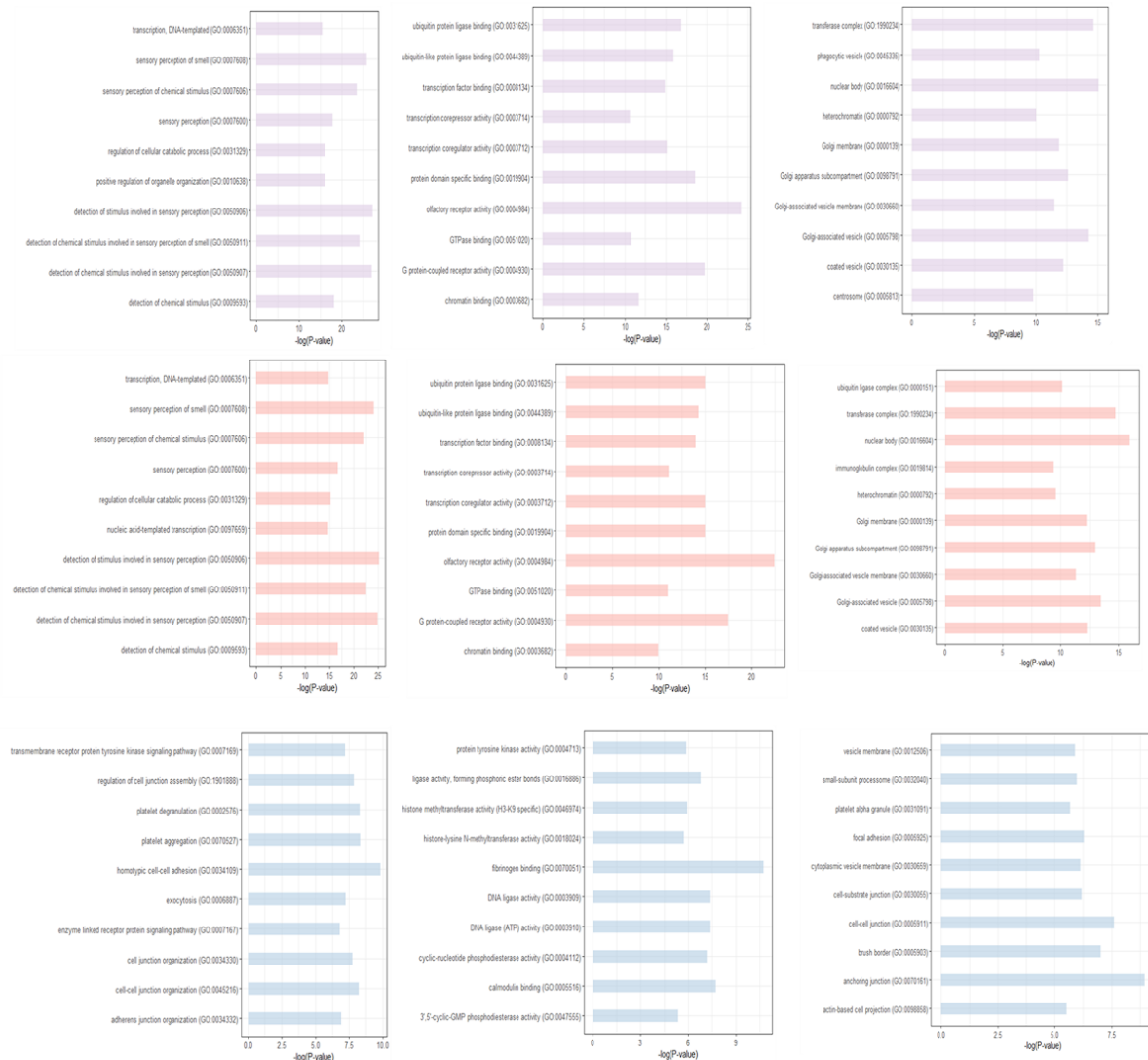### PANTHER Enrichment of Personally Perturbed Genes

The common personally perturbed genes, male specific personally perturbed genes and female personally perturbed genes were run through the PANTHER Overrepresentation Test GO Biological Process Complete, GO Molecular Function Complete, and GO Cellular Component Complete utilizing the GO Ontology database (DOI: 10.5281/zenodo.4081749 Released 2020-10-09) (**Supplemental Data 3, Supplemental Data 4, Supplemental Data 5, Supplemental Data 6, Supplemental Data 7, Supplemental Data 8, Supplemental Data 9, Supplemental Data 10, Supplemental Data 11**) [17]. Out of the set of 1189 personally perturbed genes with no gender filter, 21 of these genes did not map to any annotations. When the male and female specific genes were examined, 18 and 2 genes did not map to any annotation, respectively. These annotations were filtered to exclude any annotation with more than 1000 genes linked to it. In this way we avoid overly ambiguous annotations. The top 10 most significant annotations for each category per pool of genes is displayed in **Figure 2**.

Figure 2

Most Significant Overrepresented PANTHER GO Biological Process, GO Molecular Function, GO Cellular Component Annotations for each pool of personally perturbed genes. The PANTHER GO Ontology database (DOI: 10.5281/zenodo.4081749 Released 2020-10-09) was used to obtain these annotations [17].

The GO annotations in **Figure 2,** reveal that across all three GO categories, for genes pooled from all patients is nearly identical to the annotations for the genes pooled from all male patients. It is worth noting the cellular locations across all patients, male patients and female patients reveal an overrepresentation of genes localized to vesicle membranes. Across molecular functions transcriptional regulation/repression appears in all patients and male patients. In female patients there is an annotation for histone methyltransferase activity. Modification of histones is a known mechanism by which transcription is regulated [18]. Ubiquitin ligase activity, a top annotation for both all patients and male patients, is interesting as histone ubiquitination is another common histone modification [19].
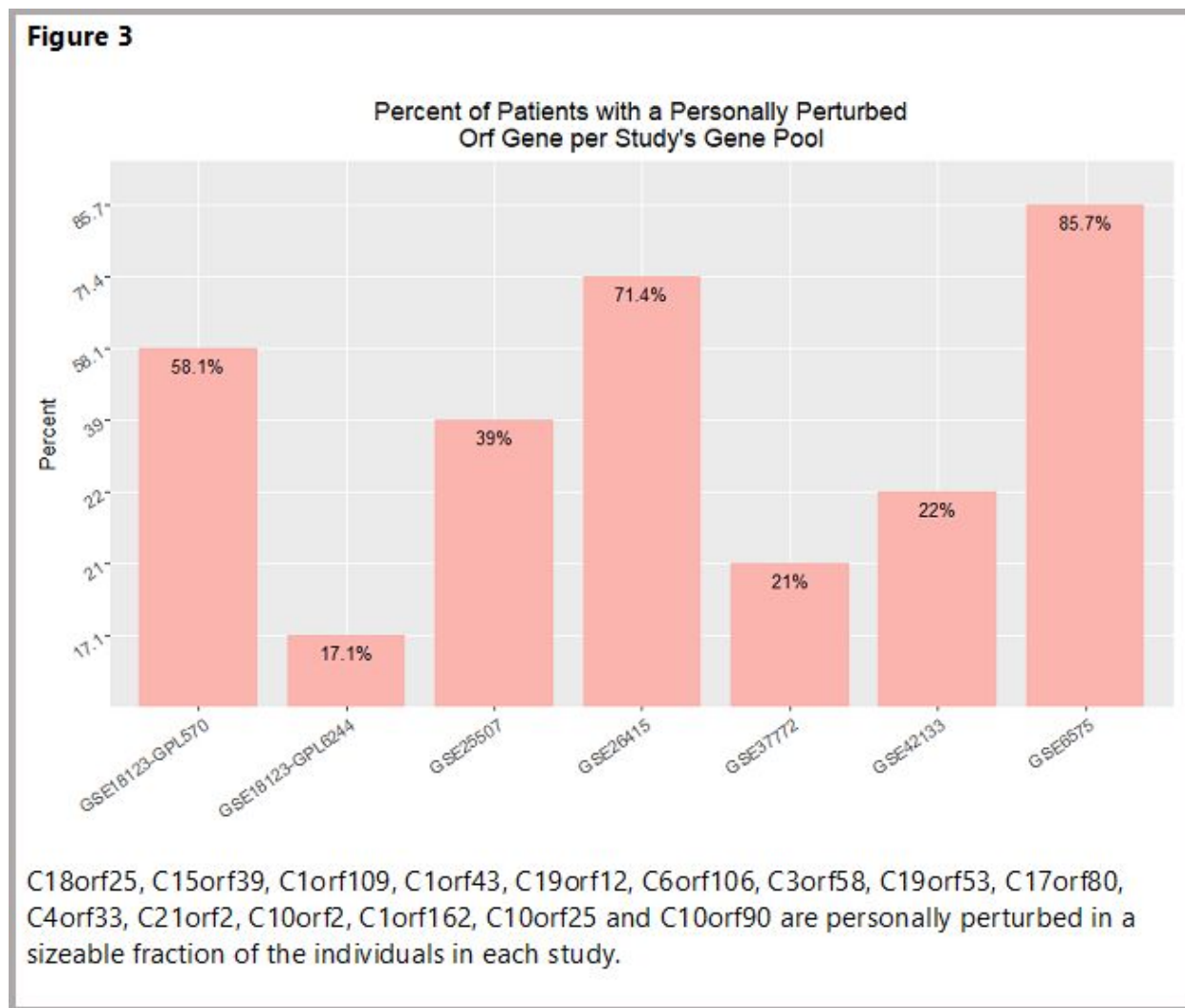
The molecular function annotations for both the pool of all and male patient genes have genes overrepresented for GTPase binding and G-protein coupled receptor (GPCR) activity. These annotations are of interest because signaling of neurotransmitters is controlled by GPCR activity [20] . The molecular function annotations for the pool of female patients show genes overrepresented for phosphodiesterase activity. Phosphodiesterases are known to break down cAMP and cGMP [21] . Intracellular signaling in motor circuits, associative/cognitive circuits and limbic circuits have been shown to be mediated by cAMP and cGMP, and as such phosphodiesterase inhibition has become an attractive drug target for neurological

disorders [22]. The presence of neurologically relevant annotations for these pools of perturbed genes suggests that the resultant genes are of some significance. The appearance of unnamed genes in these pools of perturbed genes prompted further investigation into their nature and why they might be relevant to ASD etiology.

Unnamed Genes Appear in Each Study's Pool of Personally Perturbed Genes

The gene pools for all, male and female patients contain unnamed genes. The same unnamed genes were pulled from the gene pool with no gender filter and the male gender filter; C18orf25, C15orf39, C1orf109, C1orf43, C19orf12, C6orf106, C3orf58, C19orf53, C17orf80, C4orf33, C21orf2, C10orf2, and C1orf162. The unnamed genes pulled from the female gene pool were C10orf25 and C10orf90. These genes were of interest because of their prevalence in each patient personalized perturbation profile. The prevalence of these 15 unnamed genes in each dataset ranged from 17.1% to as high as 85.7% of patients **Figure 3**. PubMed results for the male unnamed genes C18orf25, C15orf39, C1orf109, C1orf43, C19orf12, C6orf106, C3orf58, C19orf53, C17orf80, C4orf33, C21orf2, C10orf2, and C1orf162 were relatively scarce with no result count over 200 and ten of the result counts were below 20. The female unnamed genes C10orf25 and C10orf90 were also rarely mentioned in PubMed with result counts below 20 as well.

Figure 3

Percent of Patients with a Personally Perturbed Orf Gene per Study's Gene Pool

C18orf25, C15orf39, C1orf109, C1orf43, C19orf12, C6orf106, C3orf58, C19orf53, C17orf80, C4orf33, C21orf2, C10orf2, C1orf162, C10orf25 and C10orf90 are personally perturbed in a sizeable fraction of the individuals in each study.

Differential Correlation Analysis Results

Differential Gene Correlation Analysis was used to identify differentially correlated genes in ASD individuals with the unnamed genes. Unnamed genes identified in the male specific gene pool were tested for correlation against male patients in each study and the same was done for females **Figure 4**. In using differential correlation analysis, we admittingly weaken the strength of any derived correlation, since the datasets used are not only split by disease status but also by gender. It is also true small sample sizes can lead to spurious correlations [23] . However, it should also be mentioned by not separating data, spurious conclusions can also be made [23] . For instance,

Aggarwal et al. 2016 notes if hemoglobin levels are plotted against height and not separated by gender, a false correlation will appear [23] . Given the strong male bias in these data and gender-based differences in ASD presentation we thought it pertinent to query correlations after splitting by gender. By dividing by disease status, we get a unique view as to what correlations change between ASD and typically developing populations. Correlations were filtered by the ASD Spearman Correlation Coefficient greater than 0.7 or less than -0.7 and a p-value difference less than 0.05 **Figure 4**. These correlations were filtered by their presence in TCGA Pancancer Low Grade Glioma data where the Spearman Correlation Coefficient was greater than 0.3 and less than -0.3 while also having a

p-value less than 0.05 **Figure 4**. The filtered correlations were stored in **Supplemental Data 12**. Using the TCGA Pancancer Low Grade Glioma dataset as a filter was used to safeguard against spurious correlations and the ASD spearman correlation coefficient filter of 0.7 was used to ensure any correlations found were indeed strong. We use TCGA Pancancer Low Grade Glioma data not only for its size and clean data structure, but also because there is evidence of significant abnormalities in the microglia of ASD individuals leading to brain inflammation [24]. Investigation into these correlated genes was aided by the PubMed search tool Chilibot, wherein a pairwise search of gene names was used to derive connections [25].

Figure 4
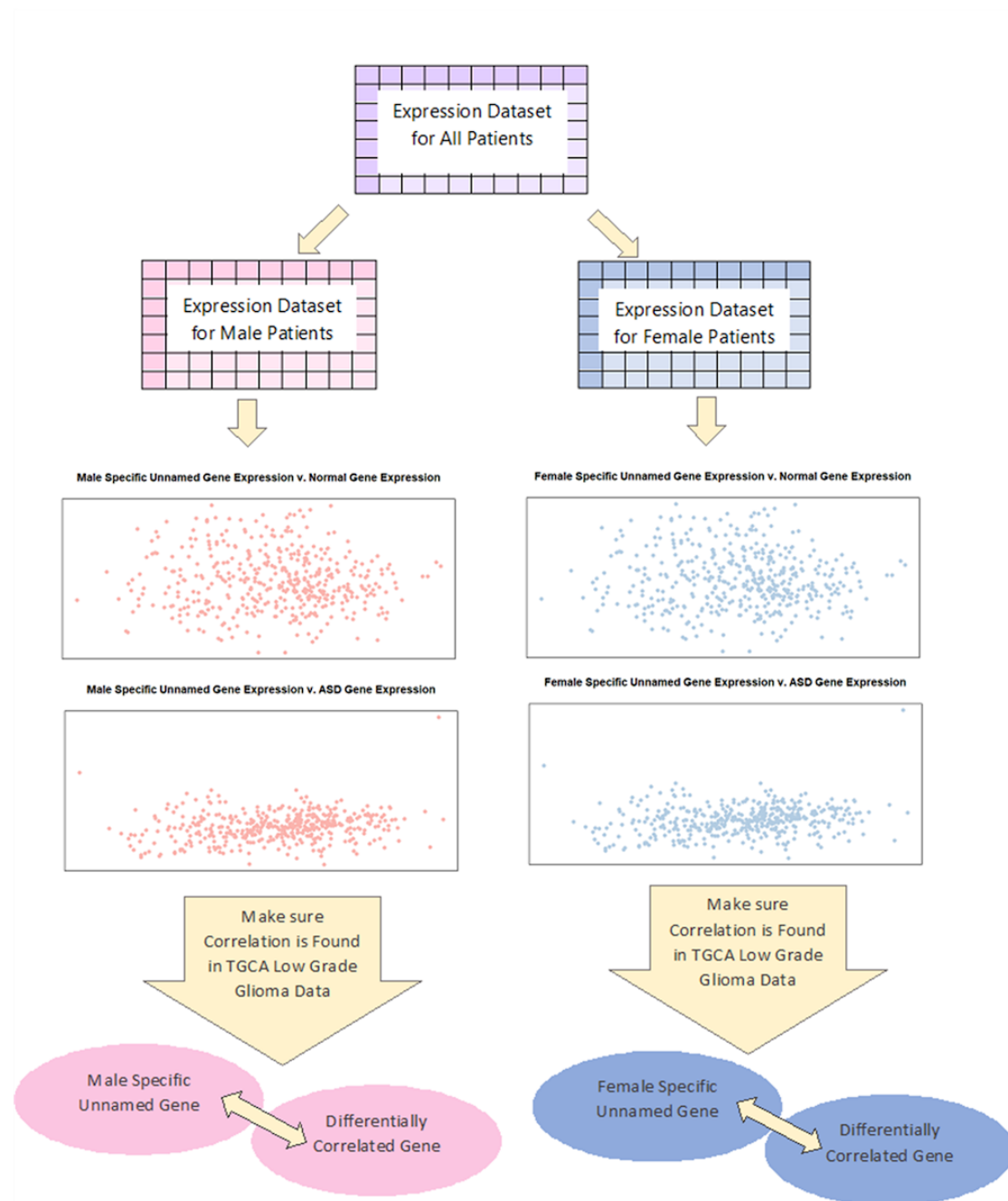
Expression Datasets were separated by gender and Differential Correlation Analysis was performed for male specific unnamed genes and female specific unnamed genes. Correlations were filtered by an ASD Spearman Correlation Coefficient greater than 0.7 and less than -0.7 while also having a p-value difference less than 0.05. These correlations were filtered by their presence in TGCA Pancancer Low Grade Glioma data where the Spearman Correlation Coefficient was greater than 0.3 and less than -0.3 while also having a p-value less than 0.05.

C18orf25, C1orf109, C6orf106, and C19orf53 have been found to localize to the nucleus [26–29]. Two of these genes, C18orf25 and C6orf106 along with C10orf90, have motifs/domains related to ubiquitin. C18orf25 has two small ubiquitin-like modifiers or SUMO motifs and C6orf106 has a ubiquitin-associated UBA-like domain [26,30]. Through the correlation analysis scheme described above we found C18orf25, C1orf109, C19orf12, C6orf106, C3orf58, C19orf53, C4orf33, C10orf2 and C10orf90 are all correlated with a variety of ubiquitin specific peptidases (USPs). USPs are the largest family of deubiquitinating enzymes, noted for their substrate specificity and association with the DNA repair pathway [31]. C10orf90 functions as an E2 ubiquitin ligase involved in responding to DNA damage [32]. C1orf109 is also implicated in DNA damage as it has been shown to lead to the accumulation of an RNA/DNA hybrid structure, called an R-loop, which has been shown to lead to DNA damage [33,34]. The helicase DHX9 was shown to also promote R-loop formation and through the correlation analysis scheme in this work we show it correlates with C10orf90, C17orf80 and C19orf53 [35]. It has also been demonstrated that C21orf2 is needed to interact with NEK1 to mitigate DNA damage repair [36]. While not directly associated with DNA damage, C17orf80 is highly correlated with CUL4, a noted E3 ubiquitin ligase, has been found to establish a DNA repair threshold [37].

C1orf43, C3orf58, and C1orf162 show localization to the Golgi apparatus while [38–40] C1orf43, C19orf12, C17orf80, C21orf2, C10orf2, have evidence of localization to the mitochondria [38,41–44]. Several of these unnamed genes have ties to phagocytosis/autophagy. By studying L. pneumophila pathogenesis , Jeng et al. 2019 stated C1orf43 as a key regulator of phagocytosis [38]. Venco et al. 2015 propose C19orf12 to be a sensor of mitochondrial damage and inducing autophagy as a protective mechanism to avoid apoptosis [45]. The variant rs3800461 on the gene C6orf106 was found by Law et al. 2017 to be implicated in autophagy [46]. Binding of C3orf58/DIPK2A to VAMP7B was shown to encourage autophagosome-lysosome fusion [47]. It is worth noting C3orf58/DIPK2A is directly implicated in familial autism [48]. Biopsies from individuals with mutations in

C10orf2 revealed autophagic vacuoles and abnormal mitochondria upon histopathological examination [49]. From the analysis performed in this work additional evidence is provided to support the connection between these poorly characterized genes and autophagy. C18orf25 is highly correlated with JAK2, HIF1A, PIK3C3, RAB10 and BMPR2 all of which are key regulators of autophagy [50–52]. C19orf53 correlates with PARK7, SOD1 and ROCK1 which have roles in autophagic proteolysis and the formation of the autophagosome [53–55]. Our correlation analysis also reveals C1orf109 is highly correlated with TMEM41B, ATG4C RAB39B, TRIM8, and NPRL3 which are also regulators of autophagy [56–59]. C1orf109 was shown here to be highly correlated with SNCA. SNCA, or alpha-synuclein, is implicated in several neurological conditions including dementia and Parkinson's disease [60]. Minakaki et al. 2018 provided evidence demonstrating components of the autophagy-lysosome pathway are present in extracellular vesicles, extracellular vesicles in cerebrospinal fluid transfer SNCA from cell to cell, and inhibiting the autophagy-lysosomal pathway increases SNCA in neuronal extracellular vesicles [61].

As mentioned earlier, C19orf12 and C10orf2 have been implicated in mitochondrial function/ disorders [45,49]. Mutations in C19orf12 have made this protein unable to localize to the mitochondrial membrane, causing high levels of mitochondrial Ca^{2+}, and an inability to respond to oxidative stress [45]. Mutations in C19orf12 are also shown to be involved with neurodegeneration with brain iron accumulation, a molecular disorder shown to be involved in the disease etiology of frontotemporal dementia, Parkinson's disease, Alzheimer's disease, Friedrich ataxia and Amyotrophic lateral sclerosis (ALS) [62]. C10orf2/TWNK is the only mitochondrial helicase needed for mitochondrial DNA replication and its dysfunction is implicated in neurodegeneration and premature aging [63]. As with phagocytosis our correlation analysis scheme has revealed interesting correlations between other unnamed genes and the mitochondria. C1orf43 correlates with VDAC1, TOMM20, and TMCO1 whose dysfunction/overexpression has been linked to mitophagy and mitochondrial dysfunction in relation to Alzheimer's disease and Parkinson's Disease [64–66]. It is

worth mentioning the two unnamed genes with direct connections to mitochondrial dysfunction, C19orf12 and C10orf2, are correlated with TMCO1 and TOMM20, respectively.

Discussion

To the best of our knowledge this is the first-time personalized perturbation profiles, as laid out by Menche et al. 2017, have been used to establish concordance among blood based ASD transcriptomic datasets. Using the Welch two sample t-test we saw only one gene, SMARCA2, was differentially expressed between all datasets. By leveraging personalized perturbation profiles, we find each study's pool of perturbed genes, male perturbed genes and female perturbed genes have 1,189, 1,111, and 133 genes in common, respectively. PANTHER GO annotations of these gene pools reveal interesting connections to vesicle membranes, GPCRs, and phosphodiesterase activity. Extracellular vesicles have recently been implicated in the disease etiology of ASD while GPCRs/phosphodiesterases are an important facet of neural signaling [20,21,67].

In these pools of perturbed genes, we found 15 unnamed genes which had a relatively high prevalence among each study's pool of perturbed genes. Of these, 13 unnamed genes were common to pools of each study's male pool of perturbed genes and two of these were common to each study's female pool of perturbed genes - the discrepancy likely owing to the smaller number of female patients. Investigating these poorly characterized genes along with differential correlation analysis demonstrated overlap between their correlations and experimentally derived functions. We find these unnamed genes have connections with DNA damage, ubiquitination, autophagy, and the mitochondria.

Mitochondrial dysfunction has been implicated in ASD. Magnetic Resonance Spectroscopy has shown N-acetyl-aspartate, a mitochondrial dysfunction marker, was significantly lower in ASD children [68]. Aberrations in respiratory chain complexes have also been found in ASD individuals [69]. Similarly, elevated levels of reactive oxygen species were observed in the brains of ASD subjects [70]. Extracellular vesicles in ASD microglia were shown to have significantly more mitochondrial DNA than typically developing individuals [67]. A recent study by Varga et al. 2018 found over 16% of ASD individuals tested in their study had deletions of mitochondrial DNA [71]. Mitochondrial mutations may also provide an answer as to why ASD seems to affect more males than females. Frank 2012 presented evidence that mitochondrial mutations act as a sex-biased sieve allowing males to build up more mitochondrial mutations [72]. Support for this theory is provided by a study using *Drosophila Melanogaster*, in which mitochondrial mutations disproportionally affect male aging as opposed to female aging [73]. Siddiqui et al. 2016 acknowledges while mitochondrial dysfunction is not yet designated a causal factor for ASD, indirect evidence for this hypothesis is rising [69].

Oxidative stress has also been related to DNA damage in relation to neurological disorders. Melnyk et al. 2012 found evidence of depleted antioxidant levels and oxidative DNA damage in the plasma of ASD individuals [74]. Markkanen et al. 2016 remarked that studies are starting to implicate DNA repair mechanisms in ASD but at times have provided conflicting results [75]. In our results we noticed C1orf109 is implicated in DNA damage through the accumulation of R-loops and C10orf90, C17orf80 and C19orf53 were highly correlated with DHX9, a protein known to promote R-loop formation [35]. In the context of neurological diseases, embryonic neural R-loop levels were determined to be essential for proper nervous system development [76]. Defective R-loop mechanisms have also been implicated in the molecular workings of ALS [77]. What is more, is evidence provided by Akman et al. 2016 show mutations in RNAase H1 have proved detrimental to mitochondrial R-loop levels, which have led to mitochondrial DNA aggregates [78]. Holt 2019 remarks this atypical organization of mitochondrial DNA is implicated in infantile-onset epilepsy, mitochondrial encephalopathy, and cerebellar dysfunction [79]. This finding might prove a useful connection between R-loops, DNA damage, mitochondrial dysfunction, and the presence of mitochondrial DNA in extracellular vesicles.

Autophagy is also starting to have a more defined role in the molecular mechanics of ASD. Microglia with defunct autophagic pathways have been shown to negatively affect synaptic pruning in ASD individuals [80]. Recently, Dana et al. 2020 found the autophagic markers LC3 and Beclin-1 were significantly disturbed in the Cc2d1a animal model of ASD [81] . What is interesting about this study is their separation by gender revealed while both ASD males and females had lower levels of Beclin-1, LC3 expression was increased in ASD females and decreased in ASD males [81] . This finding seems indicative of gender specific autophagic responses. Kasherman et al. 2020 detailed deficits in the mTOR signaling, a method by which autophagy is regulated, is disturbed in ASD subjects, and collected evidence also linking mutations in USPs to ASD features [82–84]. Kasherman et al. 2020 also states the role of USPs in ASD etiology remains poorly studied [82]. A possible answer might be found in Klusmann et al. 2018, as they provided an interesting connection between ubiquitination and R-loops, being ubiquitination of histones aided in the suppression of R-loops [85] .

More experimental validation must be done to provide conclusive links between DNA damage, R-loops, autophagy, mitochondrial damage, and ubiquitination in the context of ASD; our research here only draws attention to a handful of genes that may have been overlooked as they are poorly studied, and in the case of R-Loops, represent a relatively new field of study. The inclusion of more females in ASD blood transcriptomics would be of enormous value as well. As stated earlier, these data are derived from over 90% male ASD subjects and as such our analysis of the ASD female blood transcriptome is limited. However, we find that useful information can still be gleaned from these data with the use of personalized perturbation profiles. Aside from collecting a common pool of perturbed genes, which the t-test could not deliver, we found poorly characterized genes in those pools which may underlie ASD molecular mechanics. These results are a promising step in cataloging the true heterogeneity of ASD.

Materials and Methods

ASD Blood-Based GEO Datasets

To obtain relevant datasets we queried GEO with the terms "ASD" or "autism". These results were limited to human samples and only those of blood-based tissue types [86] . We filtered these results further by only examining studies with more than 20 ASD individuals. Expression, patient, and probe data for each study was obtained via the R package GEOquery [87]. If the expression data was not already normalized the data was normalized by adding the minimum value of the expression dataset , taking the log2 of these data and normalizing by quantiles by use of the R package limma [88] .

Personalized Perturbation Profiles and T-test

Each expression dataset was split by disease status and subjects who were not labeled as ASD were excluded from analysis. The t-test p-value was calculated using the Welch two sample t-test for each gene, with the expression of the control group being compared against the expression of the ASD group. Genes with a p-value below 0.05 were added to each study's list of differentially expressed genes. All these studies were compared with one another to obtain common differentially expressed genes. Establishing personalized perturbation profiles began with obtaining the mean and standard deviation of the expression of each gene in the control group. The z-score, for each ASD individual at each gene was obtained by the following equation as laid out by Menche et al. 2017[13] :

$$z_{ASD\_patient\_i\_gene\_X} = \frac{Exp_{ASD\_patient\_i\_gene\_X} - \mu_{control\_gene\_X}}{\sigma_{control\_gene\_X}}$$

Where, $z_{ASD\_patient\_i\_gene\_X}$ is the z-score for ASD patient i at gene X, $Exp_{ASD\_patient\_i\_gene\_X}$ is the expression of ASD patient i at gene X, $\mu_{control\_gene\_X}$ is the mean expression for the control group at gene X, and $\sigma_{control\_gene\_X}$ is the standard deviation of the expression of the control group at gene X. A gene was considered perturbed if the z-score was greater than 2.5 or less than -2.5. In this fashion each patient had a list of genes which were determined personally perturbed. These genes were added to each study's

pool of personally perturbed genes. These pools of perturbed genes were compared with one another to obtain a pool of commonly perturbed genes. The common male personally perturbed gene pool was obtained by limiting each study's pool of personally perturbed genes to male patients, and these male only study pools were compared with one another. Common female specific personally perturbed genes were obtained in the same way.

PANTHER Enrichment

Pools of personally perturbed genes were run through the PANTHER Overrepresentation Test GO Biological Process Complete, GO Molecular Function Complete, and GO Cellular Component Complete utilizing the GO Ontology database DOI: 10.5281/zenodo.4081749 Released 2020-10-09 [17]. The resulting data tables were downloaded and were filtered to exclude any annotation with more than 1000 genes linked to it. The negative log was taken of the raw p-value and only the top 10 most significant annotations were displayed. Plots were generated using the R package ggplot2 and colors were obtained using the R package RColorBrewer [89,90]. These packages were also used when plotting the prevalence of unnamed genes in each study's pool of perturbed genes.

Differential Correlation Analysis and Chilibot

Differential correlation analysis was performed using the R package DGCA [91] . Design matrices were created to split data based on ASD status. The spearman correlation coefficient used as a measure of differential correlation. Differential correlation was only obtained for the unnamed genes found in the pools of commonly perturbed genes. Only those genes with a p-value difference less than 0.05 and an ASD spearman correlation coefficient greater than 0.7 and less than -0.7 were explored further. As a level of noise reduction TCGA Pancancer Low Grade Glioma Coexpression data was used to verify correlations found by differential correlation analysis. Coexpression data mRNA Expression, RSEM Batch normalized from Illumina HiSeq_RNASeqV2, for the unnamed genes was downloaded from cBioPortal [29,92–101] . These coexpression data were filtered to exclude correlations

whose p-value is above 0.05 and whose spearman correlation coefficient is below 0.3 and above -0.3 but still below 0. Those correlations not found in these refined TCGA Pancancer Low Grade Glioma data were excluded from further analysis. The resulting correlated genes were investigated using the PubMed text mining tool, Chilibot [25]. Chilibot's pairwise search function was used to determine how correlated genes might be connected.

References

1. Kuwano, Y. et al. Autism-associated gene expression in peripheral leucocytes commonly observed between subjects with autism and healthy women having autistic children. PLoS One 6, e24723 (2011).
2. Gregg, J. P. et al. Gene expression changes in children with autism. Genomics 91, 22–29 (2008).
3. Luo, R. et al. Genome-wide transcriptome profiling reveals the functional impact of rare de novo and recurrent CNVs in autism spectrum disorders. Am. J. Hum. Genet. 91, 38–55 (2012).
4. Kong, S. W. et al. Characteristics and predictive value of blood transcriptome signature in males with autism spectrum disorders. PLoS One 7, e49475 (2012).
5. Pramparo, T. et al. Cell cycle networks link gene expression dysregulation, mutation, and brain maldevelopment in autistic toddlers. Mol. Syst. Biol. 11, 841 (2015).
6. Alter, M. D. et al. Autism and increased paternal age related changes in global levels of gene expression regulation. PLoS One 6, e16715 (2011).
7. Abrahams, B. S. et al. SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). Mol. Autism 4, 36 (2013).
8. Loomes, R., Hull, L. & Mandy, W. P. L. What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis. J. Am. Acad. Child Adolesc. Psychiatry 56, 466–474 (2017).
9. Hashem, S. et al. Genetics of structural and functional brain changes in autism spectrum disorder. Transl. Psychiatry 10, 229 (2020).

10. Ratto, A. B. *et al.* What About the Girls? Sex-Based Differences in Autistic Traits and Adaptive Skills. *J. Autism Dev. Disord.* **48**, 1698–1711 (2018).

11. Gregg, J. P. *et al.* Gene expression changes in children with autism. *Genomics* **91**, 22–29 (2008).

12. Luo, R. *et al.* Genome-wide transcriptome profiling reveals the functional impact of rare de novo and recurrent CNVs in autism spectrum disorders. *Am. J. Hum. Genet.* **91**, 38–55 (2012).

13. Menche, J. *et al.* Integrating personalized gene expression profiles into predictive disease-associated gene pools. *NPJ Syst Biol Appl* **3**, 10 (2017).

14. Balázsi, G., van Oudenaarden, A. & Collins, J. J. Cellular decision making and biological noise: from microbes to mammals. *Cell* **144**, 910–925 (2011).

15. Tomczak, A. *et al.* Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. *Sci. Rep.* **8**, 5115 (2018).

16. He, Y., Zhou, Y., Ma, W. & Wang, J. An integrated transcriptomic analysis of autism spectrum disorder. *Sci. Rep.* **9**, 11818 (2019).

17. Thomas, P. D. *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* **13**, 2129–2141 (2003).

18. Chen, J. *et al.* ER stress triggers MCP-1 expression through SET7/9-induced histone methylation in the kidneys of db/db mice. *Am. J. Physiol. Renal Physiol.* **306**, F916–25 (2014).

19. Susztak, K. Understanding the epigenetic syntax for the genetic alphabet in the kidney. *J. Am. Soc. Nephrol.* **25**, 10–17 (2014).

20. Koelle, M. R. Neurotransmitter signaling through heterotrimeric G proteins: insights from studies in C. elegans. *WormBook* **2018**, 1–52 (2018).

21. Beavo, J. A. Cyclic nucleotide phosphodiesterases: functional implications of multiple isoforms. *Physiol. Rev.* **75**, 725–748 (1995).

22. Heckman, P. R. A., Blokland, A., Bollen, E. P. P. & Prickaerts, J. Phosphodiesterase inhibition and modulation of corticostriatal and hippocampal circuits: Clinical overview and translational considerations. *Neurosci. Biobehav. Rev.* **87**, 233–254 (2018).

23. Aggarwal, R. & Ranganathan, P. Common pitfalls in statistical analysis: The use of correlation techniques. *Perspect. Clin. Res.* **7**, 187–190 (2016).

24. Takano, T. Role of Microglia in Autism: Recent Advances. *Dev. Neurosci.* **37**, 195–202 (2015).

25. Chen, H. & Sharp, B. M. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics* **5**, 147 (2004).

26. Sun, H., Liu, Y. & Hunter, T. Multiple Arkadia/RNF111 structures coordinate its Polycomb body association and transcriptional control. *Mol. Cell. Biol.* **34**, 2981–2995 (2014).

27. Liu, S.-S. *et al.* Identification and characterization of a novel gene, c1orf109, encoding a CK2 substrate that is involved in cancer cell proliferation. *J. Biomed. Sci.* **19**, 49 (2012).

28. Ambrose, R. L., Liu, Y. C., Adams, T. E., Bean, A. G. D. & Stewart, C. R. C6orf106 is a novel inhibitor of the interferon-regulatory factor 3-dependent innate antiviral response. *J. Biol. Chem.* **293**, 10561–10573 (2018).

29. Kong, W. *et al.* Diversified Application of Barcoded PLATO (PLATO-BC) Platform for Identification of Protein Interactions. *Genomics Proteomics Bioinformatics* **17**, 319–331 (2019).

30. Ambrose, R. L. *et al.* Molecular characterisation of ILRUN, a novel inhibitor of proinflammatory and antimicrobial cytokines. *Heliyon* **6**, e04115 (2020).

31. Clague, M. J. *et al.* Deubiquitylases from genes to organism. *Physiol. Rev.* **93**, 1289–1315 (2013).

32. Yan, S. *et al.* FATS is an E2-independent ubiquitin ligase that stabilizes p53 and promotes its activation in response to DNA damage. *Oncogene* **33**, 5424–5433 (2014).

33. Amon, J. D. & Koshland, D. RNase H enables efficient repair of R-loop induced DNA damage. *Elife* **5**, (2016).

34. Dou, P. *et al.* C1orf109L binding DHX9 promotes DNA damage depended on the R-loop accumulation and enhances camptothecin chemosensitivity. *Cell Prolif.* **53**, e12875 (2020).

35. Chakraborty, P., Huang, J. T. J. & Hiom, K. DHX9 helicase promotes R-loop formation in cells with impaired RNA splicing. *Nat. Commun.* **9**, 4346 (2018).

36. Fang, X. *et al.* The NEK1 interactor, C21ORF2, is

required for efficient DNA damage repair. *Acta Biochim. Biophys. Sin.* **47**, 834–841 (2015).

37. Liu, L. *et al.* CUL4A abrogation augments DNA damage response and protection against skin carcinogenesis. *Mol. Cell* **34**, 451–460 (2009).

38. Jeng, E. E. *et al.* Systematic Identification of Host Cell Regulators of Legionella pneumophila Pathogenesis Using a Genome-wide CRISPR Screen. *Cell Host Microbe* **26**, 551–563.e6 (2019).

39. Dudkiewicz, M., Lenart, A. & Pawłowski, K. A novel predicted calcium-regulated kinase family implicated in neurological disorders. *PLoS One* **8**, e66427 (2013).

40. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

41. Hinarejos, I., Machuca-Arellano, C., Sancho, P. & Espinós, C. Mitochondrial Dysfunction, Oxidative Stress and Neuroinflammation in Neurodegeneration with Brain Iron Accumulation (NBIA). *Antioxidants (Basel)* **9**, (2020).

42. Antonicka, H. *et al.* A High-Density Human Mitochondrial Proximity Interaction Network. *Cell Metab.* **32**, 479–497.e9 (2020).

43. Wang, Z. *et al.* Axial Spondylometaphyseal Dysplasia Is Caused by C21orf2 Mutations. *PLoS One* **11**, e0150555 (2016).

44. Spelbrink, J. N. *et al.* Human mitochondrial DNA deletions associated with mutations in the gene encoding Twinkle, a phage T7 gene 4-like protein localized in mitochondria. *Nat. Genet.* **28**, 223–231 (2001).

45. Venco, P. *et al.* Mutations of C19orf12, coding for a transmembrane glycine zipper containing mitochondrial protein, cause mis-localization of the protein, inability to respond to oxidative stress and increased mitochondrial Ca^{2+}. *Front. Genet.* **6**, 185 (2015).

46. Law, P. J. *et al.* Genome-wide association analysis implicates dysregulation of immunity genes in chronic lymphocytic leukaemia. *Nat. Commun.* **8**, 14175 (2017).

47. Tian, X. *et al.* DIPK2A promotes STX17- and VAMP7-mediated autophagosome-lysosome fusion by binding to VAMP7B. *Autophagy* **16**, 797–810 (2020).

48. Morrow, E. M. *et al.* Identifying autism loci and

49. Park, M.-H. *et al.* Recessive C10orf2 mutations in a family with infantile-onset spinocerebellar ataxia, sensorimotor polyneuropathy, and myopathy. *Neurogenetics* **15**, 171–182 (2014).

50. You, L. *et al.* The role of STAT3 in autophagy. *Autophagy* **11**, 729–739 (2015).

51. Palmisano, N. J. *et al.* The recycling endosome protein RAB-10 promotes autophagic flux and localization of the transmembrane protein ATG-9. *Autophagy* **13**, 1742–1753 (2017).

52. Gomez-Puerto, M. C. *et al.* Autophagy contributes to BMP type 2 receptor degradation and development of pulmonary arterial hypertension. *J. Pathol.* **249**, 356–367 (2019).

53. Lee, D.-H. *et al.* PARK7 modulates autophagic proteolysis through binding to the N-terminally arginylated form of the molecular chaperone HSPA5. *Autophagy* **14**, 1870–1885 (2018).

54. Rudnick, N. D. *et al.* Distinct roles for motor neuron autophagy early and late in the SOD1G93A mouse model of ALS. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E8294–E8303 (2017).

55. Gurkar, A. U. *et al.* Identification of ROCK1 kinase as a critical regulator of Beclin1-mediated autophagy during metabolic stress. *Nat. Commun.* **4**, 2189 (2013).

56. Moretti, F. *et al.* TMEM41B is a novel regulator of autophagy and lipid mobilization. *EMBO Rep.* **19**, (2018).

57. Maruyama, T. & Noda, N. N. Autophagy-regulating protease Atg4: structure, function, regulation and inhibition. *J. Antibiot.* (2017) doi:10.1038/ja.2017.104.

58. Tang, B. L. RAB39B's role in membrane traffic, autophagy, and associated neuropathology. *J. Cell. Physiol.* (2020) doi:10.1002/jcp.29962.

59. Wei, Y., Reveal, B., Cai, W. & Lilly, M. A. The GATOR1 Complex Regulates Metabolic Homeostasis and the Response to Nutrient Stress in Drosophila melanogaster. *G3* **6**, 3859–3867 (2016).

60. International Parkinson Disease Genomics Consortium *et al.* Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide

genes by tracing recent shared ancestry. *Science* **321**, 218–223 (2008).

association studies. *Lancet* **377**, 641–649 (2011).

61. Minakaki, G. *et al.* Autophagy inhibition promotes SNCA/alpha-synuclein release and transfer via extracellular vesicles with a hybrid autophagosome-exosome-like phenotype. *Autophagy* **14**, 98–119 (2018).

62. Wang, Z.-B. *et al.* Neurodegeneration with brain iron accumulation: Insights into the mitochondria dysregulation. *Biomed. Pharmacother.* **118**, 109068 (2019).

63. Peter, B. & Falkenberg, M. TWINKLE and Other Human Mitochondrial DNA Helicases: Structure, Function and Disease. *Genes* **11**, (2020).

64. Shoshan-Barmatz, V., Nahon-Crystal, E., Shteinfer-Kuzmine, A. & Gupta, R. VDAC1, mitochondrial dysfunction, and Alzheimer's disease. *Pharmacol. Res.* **131**, 87–101 (2018).

65. Wang, X. *et al.* ER stress mediated degradation of diacylglycerol acyltransferase impairs mitochondrial functions in TMCO1 deficient cells. *Biochem. Biophys. Res. Commun.* **512**, 914–920 (2019).

66. Franco-Iborra, S. *et al.* Defective mitochondrial protein import contributes to complex I-induced mitochondrial dysfunction and neurodegeneration in Parkinson's disease. *Cell Death Dis.* **9**, 1122 (2018).

67. Tsilioni, I. & Theoharides, T. C. Extracellular vesicles are increased in the serum of children with autism spectrum disorder, contain mitochondrial DNA, and stimulate human microglia to secrete IL-1β. *J. Neuroinflammation* **15**, 239 (2018).

68. Ipser, J. C. *et al.* 1H-MRS in autism spectrum disorders: a systematic meta-analysis. *Metab. Brain Dis.* **27**, 275–287 (2012).

69. Siddiqui, M. F., Elwell, C. & Johnson, M. H. Mitochondrial Dysfunction in Autism Spectrum Disorders. *Autism Open Access* **6**, (2016).

70. Chauhan, A. *et al.* Brain region-specific deficit in mitochondrial electron transport chain complexes in children with autism. *J. Neurochem.* **117**, 209–220 (2011).

71. Varga, N. Á. *et al.* Mitochondrial dysfunction and autism: comprehensive genetic analyses of children with autism and mtDNA deletion. *Behav. Brain Funct.* **14**, 4 (2018).

72. Frank, S. A. Evolution: mitochondrial burden on male health. *Current biology: CB* vol. 22 R797–9 (2012).

73. Camus, M. F., Clancy, D. J. & Dowling, D. K. Mitochondria, maternal inheritance, and male aging. *Curr. Biol.* **22**, 1717–1721 (2012).

74. Tang, G. *et al.* Mitochondrial abnormalities in temporal lobe of autistic brain. *Neurobiol. Dis.* **54**, 349–361 (2013).

75. Markkanen, E., Meyer, U. & Dianov, G. L. DNA Damage and Repair in Schizophrenia and Autism: Implications for Cancer Comorbidity and Beyond. *Int. J. Mol. Sci.* **17**, (2016).

76. Sorrells, S. *et al.* Spliceosomal components protect embryonic neurons from R-loop-mediated DNA damage and apoptosis. *Dis. Model. Mech.* **11**, (2018).

77. Salvi, J. S. & Mekhail, K. R-loops highlight the nucleus in ALS. *Nucleus* **6**, 23–29 (2015).

78. Akman, G. *et al.* Pathological ribonuclease H1 causes R-loop depletion and aberrant DNA segregation in mitochondria. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E4276–85 (2016).

79. Holt, I. J. The mitochondrial R-loop. *Nucleic Acids Res.* **47**, 5480–5489 (2019).

80. Kim, H.-J. *et al.* Deficient autophagy in microglia impairs synaptic pruning and causes social behavioral defects. *Mol. Psychiatry* **22**, 1576–1584 (2017).

81. Dana, H. *et al.* Disregulation of Autophagy in the Transgenerational Cc2d1a Mouse Model of Autism. *Neuromolecular Med.* **22**, 239–249 (2020).

82. Kasherman, M. A., Premarathne, S., Burne, T. H. J., Wood, S. A. & Piper, M. The Ubiquitin System: a Regulatory Hub for Intellectual Disability and Autism Spectrum Disorder. *Mol. Neurobiol.* **57**, 2179–2193 (2020).

83. Johnson, B. V. *et al.* Partial Loss of USP9X Function Leads to a Male Neurodevelopmental and Behavioral Disorder Converging on Transforming Growth Factor β Signaling. *Biol. Psychiatry* **87**, 100–112 (2020).

84. Hao, Y.-H. *et al.* USP7 Acts as a Molecular Rheostat to Promote WASH-Dependent Endosomal Protein Recycling and Is Mutated in a Human Neurodevelopmental Disorder. *Mol. Cell*

59, 956–969 (2015).

85. Klusmann, I. *et al.* Chromatin modifiers Mdm2 and RNF2 prevent RNA:DNA hybrids that impair DNA replication. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E11311–E11320 (2018).

86. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* **41**, D991–5 (2013).

87. S David, P. M. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. https://www.bioconductor.org/packages/release/bioc/html/GEOquery.html (2007).

88. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

89. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. https://cran.r-project.org/web/packages/ggplot2/citation.html (2016).

90. Neuwirth, E. ColorBrewer Palettes [R package RColorBrewer version 1.1-2]. (2014).

91. McKenzie, A. T., Katsyv, I., Song, W.-M., Wang, M. & Zhang, B. DGCA: A comprehensive R package for Differential Gene Correlation Analysis. *BMC Syst. Biol.* **10**, 106 (2016).

92. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).

93. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291–304.e6 (2018).

94. Taylor, A. M. *et al.* Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell* **33**, 676–689.e3 (2018).

95. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**, 400–416.e11 (2018).

96. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337.e10 (2018).

97. Gao, Q. *et al.* Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. *Cell Rep.* **23**, 227–238.e3 (2018).

98. Bhandari, V. *et al.* Molecular landmarks of tumor hypoxia across cancer types. *Nat. Genet.* **51**, 308–318 (2019).

99. Poore, G. D. *et al.* Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579**, 567–574 (2020).

100. Ding, L. *et al.* Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell* **173**, 305–320.e10 (2018).

101. Bonneville, R. *et al.* Landscape of Microsatellite Instability Across 39 Cancer Types. *JCO Precis Oncol* **2017**, (2017).

Supplementary Data

Supplementary Data are available at BioRxiv online.

Conflict of Interest

The authors declare no competing interests.