Cross-platform transcriptomic profiling of the response to recombinant human erythropoietin

Guan Wang (g.wang2@brighton.ac.uk)
University of Brighton

Traci Kitaoka
Illumina (United States)

Ali Crawford
Illumina (United States)

Qian Mao
BGI Group (China)

Andrew Hesketh
University of Brighton

Fergus Guppy
University of Brighton

Garrett Ash
Yale University

Jason Liu
Yale University

Mark Gerstein
Yale University

Yannis Pitsiladis
University of Brighton

Research Article

Keywords: RNA biology, MGI DNBSEQ™, BeadChip, GeneChip™, Human Transcriptome Array 2.0

Posted Date: May 19th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-510750/v1

Cross-platform transcriptomic profiling of the response to recombinant human erythropoietin

Guan Wang[1,2]*, Traci Kitaoka[3], Ali Crawford[3], Qian Mao[4], Andrew Hesketh[5], Fergus M. Guppy[5,6], Garrett I. Ash[7,8], Jason Liu[9], Mark B. Gerstein[9-12], Yannis P. Pitsiladis[6]*

[1]Sport and Exercise Science and Sports Medicine Research and Enterprise Group, University of Brighton, Brighton, UK.

[2]Centre for Regenerative Medicine and Devices, University of Brighton, Brighton, UK.

[3]Illumina, San Diego, CA, USA.

[4]BGI, Shenzhen, China.

[5]School of Pharmacy and Biomolecular Sciences, University of Brighton, Brighton, UK.

[6]Centre for Stress and Age-related Disease, University of Brighton, Brighton, UK.

[7]Veterans Affairs Connecticut Healthcare System, West Haven, CT, USA.

[8]Center for Medical Informatics, Yale University, New Haven, CT, USA.

[9]Program in Computational Biology & Bioinformatics, Yale University, New Haven, CT, USA.

[10]Department of Molecular Biophysics & Biochemistry, Yale University, New Haven, CT, USA.

[11]Department of Computer Science, Yale University, New Haven, CT, USA.

[12]Department of Statistics and Data Science, Yale University, New Haven, CT, USA.

*Corresponding authors. Email: g.wang2@brighton.ac.uk; y.pitsiladis@brighton.ac.uk.

Abstract

RNA-seq has matured and become an important tool for studying RNA biology. Here we compared two RNA-seq (Illumina sequencing by synthesis and MGI DNBSEQ^{TM}) and two microarray platforms (Illumina Expression BeadChip and GeneChip^{TM} Human Transcriptome Array 2.0) in healthy individuals administered recombinant human erythropoietin for transcriptome-wide quantification of differential gene expression. The results show that total RNA sequencing combined with DNB-seq produced a multitude of genes of biological relevance and significance in response to recombinant human erythropoietin, in contrast to other platforms. Through data triangulation linking genes to functions, genes representing the processes of erythropoiesis as well as non-erythropoietic functions of erythropoietin were unveiled. This study provides a knowledge base of genes characterising the responses to recombinant human erythropoietin through cross-platform comparison and validation.

Introduction

High-throughput technologies in gene discovery, quantification and functional investigation have advanced our understanding of complex traits and facilitated disease diagnosis, prevention and treatment over the past decade[1, 2]. Although technologies continue to evolve for discerning and characterising genes and gene-protein interactions both *ex-* and *in-vivo*, uncovering coding transcriptomes of bulk cells can capture global gene expression patterns that may directly pinpoint important biological processes at the molecular level. Which tool to use will ultimately depend on the fundamental research question. Here, we performed RNA-seq and microarray analyses in healthy individuals administered recombinant human erythropoietin (rHuEPO) to assess their discriminatory capacity, and importantly, to explore the implications of the findings to better understanding the systemic responses to rHuEPO; a first of its kind in the investigation of transcriptome-wide responses to rHuEPO in humans. This study primarily differs from previous cross-platform gene-expression studies in 1) systematic comparisons between two RNA-seq platforms (MGI DNBSEQ-G400RS and Illumina NextSeq 500), 2) comparisons with the benchmarking microarrays (GeneChip^{TM} Human Transcriptome Array 2.0 and Illumina HumanHT-12 v4 Expression BeadChip), 3) the use of a relatively large number of the same experimental samples across all four platforms, and 4) the adoption of a data triangulation approach across platforms to prioritise the functional genes of diagnostic potential.

Eighteen endurance-trained Caucasian males at sea level (Glasgow, Scotland; age: 26.0±4.5 yrs, weight: 74.8±7.9kg, height: 179.8±5.4cm) underwent 4 weeks of rHuEPO injections (50 IU/kg every 2 days)[3]. Whole blood samples collected from the 18 subjects across 8 time points — before (−14- and −1-day prior to the first injection; Base1 and Base2), during (2-, 14- and 28-day into the administration; EPO3, EPO4 and EPO5) and post rHuEPO administration (2-, 14- and 28-day after the last injection; Post6, Post7 and Post8) — were analysed on the Illumina HumanHT-12 v4 Expression BeadChip previously[4]. In the current study, 50 samples from 10 of the 18 subjects collected at Base1, Base2, EPO3, EPO4 and Post7 were analysed on the two RNA-seq platforms as well as on the GeneChip^{TM} array for quantifying differential gene expression (DGE). This experimental design aimed to identify the gene expression response to rHuEPO through robust quantification processes, and to generate results with wide applications ranging from developing effective therapeutics targeting clinical disorders associated with EPO dysfunctions to facilitating sensitive testing strategies against blood doping in sport.

Results

Total RNA DNB-seq (MGI) identifies a wealth of mRNA genes in response to rHuEpo

We identified 16,738 genes (MGI RNA-seq), 16,581 genes (Illumina RNA-seq), 29,517 transcript clusters (GeneChip), and 10,622 transcripts (BeadChip) for the DGE analyses (Table 1). Both MGI and Illumina RNA-seq generated good base call quality, with an average quality score of >34 across the read lengths and across the samples (Supplementary Fig. 1). No sample contamination/swaps (Supplementary Fig. 2) and no other significant surrogate variables of batch effects were detected in these sequencing datasets. Genome mapping using HISAT2[5] (against the reference genome assembly GRCh38.p12[6]) showed the overall alignment rates of 94.2% (MGI; 197.9M total reads) and 95.0% (Illumina; 110.4M total reads) (Supplementary Table 1). RseQC[7] revealed a large proportion of the sequences aligned to introns in the MGI RNA-seq data (37.4% versus 8.7% Illumina on average; Supplementary Fig. 3 and Supplementary Table 2), a result coinciding with the differing sequencing library preparation methods used (total RNA-seq with rRNA depletion and globin mRNA reduction, MGI versus mRNA enrichment, Illumina). RseQC also showed that among a total of 186.6M (MGI) and 104.6M (Illumina) averaged reads observed, 52.4% (~ 97.8M reads) and 74.8% (~ 78.2M reads) of the reads were effectively mapped to the coding sequences (exons), respectively (Supplementary Table 2). The average Salmon[8] transcriptome mapping rates, following selective-alignment-based lightweight mapping, were 38.8% (38.3M aligned reads; MGI) and 81.9% (45.1M aligned reads; Illumina) (Supplementary Table 3). The seeming discrepancies observed in total reads and alignment rates across the software tools (HISAT2, RseQC and Salmon) were expected given their specific usage. Overall, these data suggested high quality sequences obtained from both sequencing platforms. For the purposes of cross-platform comparison, the relative abundance estimates of transcripts after Salmon transcriptome mapping were summed to gene level, and genes were considered expressed when the gene-level abundance estimates were equal to or more than 5 in at least 4 samples; resulting in the exclusion of 17,198 and 18,347 genes from the MGI and Illumina RNA-seq datasets, respectively (Table 2). Gene annotation resulted in 3,852 (MGI) and 2,860 (Illumina) un-defined gene mappings removed from the sequencing datasets (Table 2). As a result, 16,738 and 16,581 protein-coding genes identified from MGI and Illumina sequencing, respectively, were used for the downstream DGE analyses (Table 1 and 2).

Initial quality control metrics revealed variability in eight out of the fifty GeneChip^{TM} arrays (Supplementary Fig. 4A). Two of the eight samples were then repeated for chip scanning, and the other six samples were repeated from the target preparation step (Supplementary Fig. 4B). Raw intensity values obtained from the GeneChip and BeadChip analyses correspond to 67,480 and 47,286 coding and non-coding transcriptomic features, respectively (Table 1). The process of normalisation and filtering unveiled 29,517 transcript clusters (GeneChip) and 10,622 transcripts (BeadChip) as identified features (Table 1), with the detailed filtering steps and the resulting number of features summarised in Table 2. Briefly, 18,494 and 6,900 probes were removed as undetected and low-quality probes, respectively, from the BeadChip dataset (Table 2). While 8,166 probes were removed due to low average expression (cutoff value: 5.1) in the BeadChip dataset, no such probes were necessarily excluded from the GeneChip dataset (Supplementary Fig. 5 and Table 2). No significant surrogate variables representing the underlying biases, potentially arising from library preparation and/or scanning, thereby confounding the biological effects being studied, were observed in the two microarray datasets.

Unsupervised principal component analysis (PCA) revealed substantial variance, estimated using the top 500 genes ranked by expression variance across all samples. Variances explained by the principal component 1 and the principal component 2 were: 69% vs. 5% (MGI RNA-seq), 44% vs. 9% (Illumina RNA-seq), 58% vs. 14% (GeneChip), and 78% vs. 7% (Beadchip) (Supplementary Fig. 6). Gene clustering of the top 30 genes of high variance showed a good distinction across biological conditions in all datasets (Supplementary Fig. 7). Nevertheless, a more distinctive pattern across the conditions was observed following MGI RNA-seq compared to Illumina RNA-seq and GeneChip (Supplementary Fig. 7, A versus B, C). In contrast with the discrimination pattern presented in the 143 BeadChip samples, a higher expression level of the examined top 30 genes was detected by MGI RNA-seq (Supplementary Fig. 7, A versus D). The DESeq2[9] and limma[10] DGE analyses yielded 1,552, 582, 252 and 2,372 transcriptomic features exceeding the pre-defined thresholds following MGI RNA-seq, Illumina RNA-seq, GeneChip^{TM}

and BeadChip, respectively (thresholds for RNA-seq: a fold change of 1.2 and *s*-value of 0.005; for microarray: a fold change of 1.2 and BH adjusted *p*-value of 0.05; note that the probability thresholds bound to the fold change of 1.2) (Table 3). A significant proportion of these findings were unique to MGI RNA-seq at EPO4 (66.8%) and Post7 (54.5%) (Supplementary Table 4). Notably, substantial sub-proportions of the gene features identified from MGI RNA-seq exceeded an absolute fold change of 2 (12.4% at EPO4 and 18.0% at Post7) and captured even smaller changes between 1.2 and 2 (54.4% at EPO4 and 36.5% at Post7), when compared to the Illumina RNA-seq and GeneChip^{TM} gene features (ranging from 0% to 19.4%; Supplementary Table 4 and Fig. 1). Furthermore, strong correlations between the two RNA-seq platforms on the commonly identified genes were observed ($r = 0.74$ at EPO4 and $r = 0.85$ at Post7, $P < 2E-16$; Fig. 2, a and d), whereas the correlations ranged from very weak ($r = 0.2$) to moderate ($r = 0.7$) when compared RNA-seq to GeneChip^{TM} ($P < 0.0003$; Fig. 2, b, c, e and f). Overall, MGI RNA-seq, the total RNA DNB-seq, resulted in an increased sensitivity in identifying coding genes in response to EPO compared to the Illumina mRNA-seq and GeneChip^{TM} (Fig. 1, Fig. 2 and Supplementary Data 1).

Pathway analysis links the differentially expressed genes to erythropoiesis and non-erythropoietic functions of EPO

To explore the biological functions of the gene features identified from sequencing and microarray, we performed a standard GSEA run (v4.0.3) subject to 1,000 phenotype permutations[11, 12] on all datasets, using the MSigDB (v7.2)[11, 13] hallmark (H)[14] and Gene Ontology (C5; BP: GO biological process)[15, 16] collections of functional gene sets. As expected, heme metabolism emerged as the most significantly enriched pathway in all datasets following the analysis on the 50 hallmark gene sets (FDR: MGI = 0.011 EPO4, Illumina = 0.033 EPO4, GeneChip ≤ 0.017 EPO4/Post7 and BeadChip ≤ 0.004 EPO3/4/5/Post7/8; Supplementary Table 5). Leading edge genes, those contributing the most to the enrichment score of the heme metabolism pathway constituting 200 genes, included 144 (MGI; EPO4), 105 (Illumina; EPO4), 125/96 (GeneChip; EPO4/Post7) and 101/103/103/97/84 (BeadChip; EPO3/4/5/Post7/8) genes found in these datasets (Supplementary Data 2). Fifty-six leading edge genes overlapped across all platforms and across conditions (pathway FDR < 0.1) (Supplementary Data 2). Of the 56 genes, 51 and 34 genes were also identified by the standard DGE analyses for the EPO4 and Post7 conditions, respectively, across two or three of the MGI RNA-seq, Illumina RNA-seq and GeneChip^{TM} platforms (Supplementary Data 3). GSEA was able to detect the associated genes that have fallen off the detection thresholds in the standard DGE analyses of Illumina RNA-seq and GeneChip datasets (Supplementary Data 3). In addition, subsets of 10, 51, 51, 36, and 19 of the 56 leading edge genes were found in the BeadChip DGE results across EPO3, EPO4, EPO5, Post7 and Post8 conditions, respectively (Supplementary Data 4). The data suggest the effectiveness of all four detection platforms and the effectiveness of GSEA in capturing the most context-relevant biological pathway in response to rHuEPO. Next, GSEA was conducted on 7,530 GO biological processes included in the MSigDB C5 collection, and identified a total of 212, 134, and 33 biological pathways from MGI RNA-seq (EPO4), GeneChip (EPO4) and BeadChip (EPO4, EPO5, Post7 and Post8) datasets, respectively, exceeding the pathway FDR < 0.1 and nominal $P < 0.05$. No significantly enriched GO biological processes were identified from GSEA in the Illumina RNA-seq datasets. From the MGI RNA-seq dataset, these included biological processes, resembling EPO cytoprotective functions and the downstream signal transduction pathways[17-19], typically involved in response to oxidative stress (e.g. positive regulation of mitophagy, hydrogen peroxide metabolic process, and nucleotide-excision repair, DNA damage recognition), heme formation (e.g. porphyrin-containing compound metabolic process), erythrocyte development, mTOR (target of rapamycin) signaling, regulation of energy metabolism (e.g. regulation of generation of precursor metabolites and energy), low density lipoprotein clearance, and nervous system development (Fig. 3). Key pathways characterising the responses to EPO, such as autophagy of mitochondrion, positive regulation of cell cycle arrest, iron ion homeostasis, tetrapyrrole metabolic process, erythrocyte development, and ventricular system development also were identified from the GeneChip dataset (Supplementary Fig. 8). In addition, other biological processes, including cyclic GMP mediated signaling, positive regulation of cardiac muscle cell proliferation, and gamma-aminobutyric acid transport, were observed, to name a few (Supplementary Fig. 8). In the BeadChip datasets, particular pathways identified that were common to those observed on both the MGI RNA-seq and the GeneChip platforms included hemoglobin metabolic process, erythrocyte development, and hydrogen peroxide metabolic process (Supplementary Fig. 9). Further pathways of negative regulation of necrotic cell death, negative regulation of TORC1 signaling, cellular response to monoamine stimulus, monoamine transport, gas transport, lipid transport, drug transmembrane transport, synaptic signaling, synapse organisation, and multicellular organism development, were found in the BeadChip datasets (Supplementary Fig. 9). Leading edge genes from top 34, 14, and 16 pathways defined by the normalised enrichment score (NES) > 1.90 and from 38, 66, and 12 the most enriched pathways representing a biological theme where NES < 1.90 were further investigated in the MGI RNA-seq, GeneChip and BeadChip datasets, respectively. Out of a total of 308 leading edge genes identified in the MGI RNA-seq EPO4 dataset overlapping with the DESeq2 EPO4 DGE results, 135 genes also were found to be significantly expressed in the Post7 condition following the DESeq2 analysis (Supplementary Data 5). Of the 135 genes, top 10 genes filtered based on the GSEA ranks and pathway NESs — *BPGM*, *ALAS2*, *PKD1L3*, *SLC4A1*, *AP2A1*, *IGF2*, *FAM210B*, *DYRK3*, *FECH* and *SLC25A37* — characterise erythrocyte development, heme formation, metal ion homeostasis, cellular response to PH, LDL particle clearance, glucose and energy metabolism, and TOR signaling (Supplementary Data 5). Fifty-seven leading edge genes of the GeneChip EPO4 dataset were common to the limma EPO4 DGE genes, while 15 (of the 57) were also present in the Post7 DGE results (Supplementary Data 6). These 15 genes — *ALAS2*, *SLC4A1*, *FOXO3*, *TMOD1*, *FECH*, *SLC6A8*, *SLC25A39*, *SNCA*, *FAM210B*, *EPB42*, *SLC25A37*, *YBX3*, *BPGM*, *STRADB*, and *BCL2L1* — correlate with heme formation, bicarbonate transport, muscle atrophy, lens fiber cell development, gamma-aminobutyric acid transport, erythrocyte development, cellular hyperosmotic response, negative regulation of signal transduction in the absence of ligand and cellular response to amino acid stimulus (Supplementary Data 6). Among 376 leading edge genes identified from the BeadChip datasets, 76 also were observed in the limma DGE analysis. Top 10 genes (of the 76 genes) — *KCNJ10*, *YBX3*, *SNCA*, *OR2W3*, *IRX1*, *OR2W5*, *CAMK2A*, *ACP4*, *NCDN* and *HOXC10* — are involved in regulation of neuronal synaptic plasticity and necrotic cell death, sensory perception of smell, proximal/distant pattern formation, and cell fate specification, and were enriched in the GSEA Post7 dataset as well as were significantly expressed across EPO4, EPO5, Post7 and Post8 conditions following the limma DGE analysis (Supplementary Data 7). Among the above 135, 15 and 76 leading edge genes identified on the three platforms, *BPGM*, *ALAS2*, *SLC4A1*, *FAM210B*, *EPB42*, *SNCA*, *YBX3* and *TMOD1* were detected by all three platforms (Supplementary Data 8). *FECH*, *SLC25A37*, *FOXO3*, *BCL2L1*, and *SLC25A39* were common between MGI RNA-seq and GeneChip, *SLC6A8* between GeneChip and BeadChip, and *SLC7A5*, *PINK1*, *DMTN*, *TRIM58*, *SESN3*, *GATA1*, *FURIN*, *HBQ1*, *EIF2AK1*, and *HBM* between MGI RNA-seq and BeadChip (Supplementary Data 8). One hundred and twelve leading edge genes (top 5: *PKD1L3*, *AP2A1*, *DYRK3*, *IGF2* and *TAL1*) were uniquely identified by MGI RNA-seq, 1 (*STRADB*) by GeneChip and 57 by BeadChip (top 5: *KCNJ10*, *OR2W3*, *IRX1*, *OR2W5* and *CAMK2A*) (Supplementary Data 8). Further, 43 leading edge genes identified from one or more of the three platforms were detected by Illumina RNA-seq across EPO4 and Post7 conditions following the DESeq2 DGE analysis (Supplementary Data 9).

To follow up on the DGE and GSEA results, we performed additional analysis using the Reactome database to examining the pathway components inferred from the 43 genes in pathway diagrams and to confirming the gene functions attributed to rHuEPO across the experimental conditions. By overlaying the gene expression values on Reactome pathway diagrams (release 73)[20], 13 and 8 significantly expressed interaction networks represented by 29 and 13 of the 43 genes, or their interactors (IntAct score ≥ 0.556), were identified in the MGI RNA-seq and BeadChip datasets, respectively (pathway FDR < 0.05; see Supplementary Data 10 for pathway entities and statistics and Supplementary Data 11 for the corresponding pathway overviews). Notably, pathway components in the entire cascade of O_2/CO_2 exchange in erythrocytes were the most significantly altered in the MGI RNA-seq datasets as opposed to findings obtained from the Illumina RNA-seq, GeneChip and BeadChip platforms, including the pathway genes *SLC4A1*, *HBB*, *CA1*, *AQP1*, *RHAG*, *HBA1* and *CYBSR1* (pathway FDR ≤ 0.003, Supplementary Data 10) across EPO4, up-regulation and Post7, down-regulation (see Fig. 4 for the enhanced high-level pathway diagram of the Post7 dataset). Finally, by overlapping a total of 172 and 91 significantly expressed Reactome pathway genes and their interactors (IntAct score > 0.9 of high confidence interactions, Supplementary Data 10) emerged from the 13 and 8 networks with the genes identified from the standard DGE analyses in the MGI RNA-seq and BeadChip datasets, 80 and 41 genes were further confirmed, respectively (Supplementary Data 12). These 80 and 41 genes represent the candidate genes that warrant further studies to rule out potential confounding factors that mimic the EPO effect in terms of developing robust anti-doping gene signatures, or to verify the role of the genes in EPO production and function for therapeutic purposes. The subsets of the top 10 genes (sorted by the standard DGE *s*-value < 0.005 or FDR < 0.05) accompanied by their corresponding GSEA and Reactome pathways are presented in Supplementary Table 6 and 7.

Discussion

Taken together, cross-platform comparison in 10 subjects administered rHuEPO (50 IU every 2 days for 4 weeks) was conducted following gene expression quantification on MGI DNBSEQ-G400RS, Illumina NextSeq 500 and GeneChip^{TM} HTA2.0 platforms. To initiate a direct comparison, only the coding gene features were extracted and compared across platforms. There was a 2.28-fold increase in genes significantly expressed following MGI RNA-seq, as compared to the combined number of genes identified on the other two platforms (Fig. 1 and Supplementary Table 4). Furthermore, among 1,126 genes identified at EPO4, 25.5% of the genes overlapped between MGI RNA-seq and the other two platforms, and 66.8% of the genes were unique to MGI RNA-seq; among 674 genes identified at Post7, the corresponding figures were 21.5% and 54.5%, respectively (Supplementary Table 4). Among genes with an absolute fold change less than 2, Illumina RNA-seq captured a much higher proportion of the identified gene features compared to GeneChip (10.3% vs 1.7% EPO4; 26.4% vs 0.6% Post7;

Supplementary Table 4). The experimental effect of EPO was largely captured by MGI RNA-seq (PC1: 69% vs PC2: 5%), followed by GeneChip (PC1: 58% vs PC2: 14%) and Illumina RNA-seq (PC1: 44% vs PC2: 9%) by examining the top 500 genes showing the highest variability across samples (Supplementary Fig. 6, A to C). These observations support the supreme performance of total RNA DNB-seq on MGI DNBSEQ-G400RS, followed by mRNA-seq on Illumina NextSeq 500 and GeneChip^{TM} HTA2.0 in this study. Nevertheless, genes characterised by Illumina HumanHT-12 v4 Expression BeadChip in the 18 subjects represented a total of 85% of variance captured by PC1 (78%) and PC2 (7%) (Supplementary Fig. 6D), suggesting increased statistical power owing to the larger sample size (i.e. 143 samples from 18 subjects in contrast to 50 samples from 10 subjects comprising the sample sets analysed on the other platforms). Following on, quantitative pathway analysis by GSEA identified the heme metabolism pathway enriched in all datasets across the four high-throughput gene quantification platforms when analysing the MSigDB Hallmark collection of functional gene sets, while 212, 134, and 33 enriched pathways were unveiled from MGI RNA-seq, GeneChip and BeadChip datasets, respectively, by examining the MSigDB C5 collection of 7,530 biological processes. The pathway results underpinned the biological relevance of the gene expression findings, particularly with a wealth of functional information emerged from the MGI RNA-seq dataset (Fig. 3). Pathways of interest were prioritised by focusing on pathways of NES > 1.9 and representative pathways of the biological themes where the NES < 1.9. Three hundred and eight, 57 and 376 leading edge genes were extracted from the pathways of interest, eventually led to 135, 15 and 76 genes also confirmed by the standard DGE analyses of MGI RNA-seq, GeneChip and BeadChip datasets, respectively (Supplementary Data 5-7). Among these genes, 43 were further validated in the list of genes resulted from the Illumina RNA-seq DGE analysis

(Supplementary Data 9). Despite strong positive correlations observed at the gene level between MGI RNA-seq and Illumina RNA-seq, the lack of significantly expressed pathways following GSEA in the Illumina RNA-seq datasets is in line with the generally weaker signals being picked up by Illumina RNA-seq in this study (Fig. 2). To better understand the interacting networks or signaling cascades represented by the 43 genes, we explored the Reactome database and generated a total of 21 pathway overviews detailing the pathway entities/genes, their expression levels, and their interactions with other entities within the pathway or across different pathways (Supplementary Data 11). Finally, by extracting the significantly altered genes involved in these Reactome networks and by matching these genes to the results of the standard DGE analyses, we concluded with the lists of 80 and 41 genes that are of biological relevance to rHuEPO, identified on the MGI RNA-seq and BeadChip platforms, respectively (Supplementary Data 12). They represent the top biological pathways enriched in metabolism of porphyrins, O_2/CO_2 exchange in erythrocytes, response to oxidative stress induced cellular senescence, and tissue damage caused by amyloid deposition.

This comprehensive profiling of rHuEPO gene expression based on both RNA-seq and microarrays has generated a robust set of genes of biological significance in relation to erythropoiesis as well as non-erythropoietic effects of rHuEPO. It also establishes a knowledge base of genes capturing a wide range of magnitude of changes attributable to rHuEPO by RNA-seq, highlighting advantages of total RNA-seq combined with DNB-seq in quantifying gene transcription. The adoption of a data triangulation approach by cross-platform comparisons and by linking genes to their functions reinforces the biological findings and mitigates gene expression perturbations caused by normal physiological changes such as seasonal changes and lifestyle related changes. The longitudinal nature of the current investigation in healthy individuals would help facilitate detailed studies of erythroid disorders and help formulate target therapeutics, through disrupting and examining the mechanisms of the putative genes involved in erythropoiesis and non-erythropoietic functions of rHuEPO. Finally, this study underpins the follow-up studies needed to develop sensitive and robust gene signatures of blood doping in sport.

Methods

Subjects

In a previously funded research project by the World Anti-Doping Agency (grant no.: 08C19YP), we collected whole blood samples from 18 endurance-trained Caucasian males at sea level from Glasgow, Scotland (26.0±4.5 yrs, 74.8±7.9 kg, 179.8±5.4 cm), who underwent 4-week 50 $\mathrm{IU \cdot kg^{-1}}$ body mass of rHuEPO every second day[3]. Daily oral iron supplementation (100 mg of elemental iron, ferrous sulphate tablets, Almus, Barnstable, UK) was given during the 4 weeks of rHuEPO administration[3]. Whole blood samples were collected at baseline (2 weeks and 1 day before rHuEPO; denoted by Base1 and Base2, respectively), during the rHuEPO administration (2 days, 2 and 4 weeks following the 1st injection; denoted by EPO3, EPO4 and EPO5, respectively) and for 4 weeks after the rHuEPO administration (1, 2 and 4 weeks after the last injection; denoted by Post6, Post7 and Post8, respectively) for gene expression profiling on the HumanHT-12 v4.0 Expression BeadChip (Illumina, San Diego, CA, USA)[4]. In the current study (grant no.: ISF15E10YP), samples from 10 out of the 18 subjects collected at Base1, Base2, EPO3, EPO4 and Post7 were analysed on a new microarray platform (GeneChip™ Human Transcriptome Array 2.0 or HTA2.0, Thermo Fisher Scientific, Waltham, MA, USA) and on two RNA-seq platforms (NextSeq500, Illumina, San Diego, CA, USA, and DNBSEQ-G400RS, MGI Tech, Shenzhen, China) for cross-platform gene expression comparisons for robust detection of EPO gene signatures. The studies were approved by the University of Glasgow Ethics Committee (Scotland, UK) and the University of Brighton Ethics Committee (England, UK) and were performed in accordance with the "Declaration of Helsinki". Written informed consent was obtained from all subjects.

RNA collection and preparation

Three milliliters of whole blood was collected from an antecubital vein using Tempus™ Blood RNA tubes (Thermo Fisher Scientific, Waltham, MA, USA). Each Tempus™ tube contains 6 mL of RNA stabilising reagent and was vigorously mixed immediately after collection for 10 s. The blood samples were incubated at room temperature for approximately 3 hours and then stored at −20°C or −80°C before subsequent analysis or transportation to the analytical lab. Total RNA was isolated from the whole blood according to the manufacturer's instructions (Tempus™ Spin RNA Isolation Kit, Thermo Fisher Scientific, Waltham, MA, USA). The purified total RNA was eluted in 90 µL elution buffer and stored in three aliquots at −80°C until further analysis. Initial RNA quantity and purity was assessed by the Nanodrop™ ND-2000 Spectrophotometer (Thermo Fisher Scientific, Waltham, MA, US). RNA integrity was assessed using the Agilent 2100 Bioanalyser (Agilent Technologies, Santa Clara, CA, USA) prior to the RNA-seq and GeneChip analyses.

Microarray analysis with HumanHT-12 v4.0 Expression BeadChip

Detailed sample preparation for the Illumina microarray experiment are available elsewhere[4]. Briefly, 500 ng of total RNA was used for complimentary RNA (cRNA) synthesis using the Illumina™ TotalPrep RNA Amplification Kit (Thermo Fisher Scientific, Waltham, MA, USA). Seven hundred and fifty nanograms of the purified labelled cRNA samples were hybridised to the HumanHT-12 v4.0 Expression BeadChip arrays containing > 47,000 probes, following the manufacturer's recommended procedures (Illumina, San Diego, CA, USA). The Bead arrays were scanned on the Illumina BeadArray Reader. In this current study, the raw intensity values were exported using the Illumina GenomeStudio software (v2.0; Gene Expression Module). The bioconductor "limma" package[10] was used for background correction, data normalisation (using the "neqc" function)[21] and differential gene expression analysis (DGE)[22] for paired samples (using the "treat" function) in the 18 subjects across all 8 time points (i.e. Base1, Base2, EPO3, EPO4, EPO5, Post6, Post7 and Post8). Notably, only probes expressed in at least 7 samples at a detection $p<0.05$ were kept. Probes were annotated to illuminaHumanv4.db[23] and only probes with "good" and "perfect" matching quality were retained followed by removing probes with "NA" or multiple mappings. Probes with low expression values below 5.1 were excluded prior to the DGE analysis (assessed using the limma "plotSA" function). Transcripts were considered significantly expressed for a fold change of 1.2 bounded to a 5% false discovery rate (FDR) (thereby, giving more weight to fold change for gene ranking). These are common cut-off values being used for declaring biologically and statistically significant findings in a DGE analysis[24].

Microarray analysis with GeneChip™ HTA2.0

One hundred nanograms of total RNA was processed using the GeneChip™ WT Plus Reagent Kit according to the manufacturer's instructions (Thermo Fisher Scientific, Waltham, MA, US) for 10 out of the 18 subjects at the selected time points (i.e. Base1, Base2, EPO3, EPO4 and Post7). Single-stranded cDNA (ss-cDNA) was synthesised by the reverse transcription of cRNA. Two hundred microlitres of hybridisation cocktail (containing approximately 5.2 µg fragmented and labelled ss-cDNA) was loaded onto the GeneChip^{TM} HTA2.0 (Thermo Fisher Scientific, Waltham, MA, US). The GeneChip^{TM} arrays were incubated in the GeneChip^{TM} Hybridization Oven 645 for 16 hours, washed and stained on the GeneChip^{TM} Fluidics Station 450. The arrays were then scanned using the GeneChip^{TM} Scanner 3000 7G. The Applied Biosystems^{TM} Transcriptome Analysis Console (version:4.0.1.36; Thermo Fisher Scientific, Waltham, MA, US) was used to perform initial data QC and data visualisation. The relative log expression box plots were plotted following the quality assessment steps illustrated in *ref*[25]. The Bioconductor "oligo" package[26] was used to read in the raw intensity CEL files, and the "rma" function was used for background correction, normalisation, and data summarisation to the gene level (defined by the argument "core"). Probes were annotated to hta20transcriptcluster.db[27] and probes with "NA" or multiple mappings were removed. The "limma" package was then used to perform the usual DGE analysis for paired samples (the analysis setting is identical to that used in the Illumina microarray analysis illustrated above). Transcript clusters (loosely equal to genes) were considered significantly expressed at a fold change of 1.2 bounded to a 5% FDR.

RNA-seq on Illumina NextSeq500

Five hundred nanograms of total RNA was used for sequencing according to the Illumina TruSeq Stranded mRNA sample prep guide - high sample protocol (Illumina, San Diego, CA, USA). Briefly, mRNA molecules were purified using the poly-T oligo attached magnetic beads following which the mRNA was fragmented and primed for cDNA synthesis. A single "A" base was subsequently added to the 3-prime end of the synthesised blunt-ended cDNA and ligated with index adapters for hybridisation onto a flow cell. The DNA fragments with adapters on both ends were amplified via polymerase chain reaction to generate the final double-stranded cDNA (ds-cDNA) library followed by library validation and normalisation and pooling of the samples. Samples were pooled and then sequenced at 2x75 bp read length to a depth of approximately 64 M reads per sample on the Illumina NextSeq 500 (Illumina, San Diego, CA, USA). 10 out of the 18 subjects at the selected time points (i.e. Base1, Base2, EPO3, EPO4, and Post7) were analysed. Raw sequences were examined by FastQC[28] for basic quality checks (e.g. per base sequence quality, adaptor content, and per base N content), FastQ Screen[29] for mapping against multiple reference genomes for detecting sample swaps or sample contamination that may have resulted from sources other than humans (i.e. in this case, mapping against human, mouse and rat genomes were conducted), HISAT2[5] for alignment to the reference genome assembly (GRCh38.p12[6]) using the Ensembl 94 annotation[30] prior to RseQC[7] for read distribution analysis, Salmon[8] for aligning to the transcriptome and transcripts quantification (using selective alignment with the *decoy aware* target transcriptome to eliminate potential spurious mapping to unannotated genomic locus over a *k*-mer length of 31, along with --SeqBias and --gcBias flags switched on to correct for any unwanted effects), bioconductor package "tximport"[31] for summarising transcript-level estimates to genes based on Ensembl release 94[30], and DESeq2[9] for paired sample DGE analysis. Pre-filtering was performed to keep genes that have at least 5 reads across 4 samples prior to the DGE analysis. Ensembl IDs were mapped to gene symbols using the bioconductor package "org.Hs.eg.db"[32] and un-defined mappings were removed (i.e. gene with "NA" or multiple mappings). MultiQC[33] was used to aggregate the analysis results from the FastQC, FastQ Screen and RseQC runs from multiple samples.

Unsupervised principal component analysis (PCA) for top 500 genes of high variance and gene clustering analysis for the top 30 genes were performed following the DESeq2 vignette on data quality assessment procedures[34]. The bioconductor package "SVA"[35] was used to assess surrogate variables that may represent other variations in the data for further correction. Shrinkage estimator "apeglm" was used for the shrinkage of log fold change estimates and for ranking genes by effect size[36]. Genes exceeding a fold change of 1.2 bounded to the default *s*-value < 0.005 were reported.

RNA-seq on MGI DNBSEQ-G400RS

Four hundred nanograms of total RNA was used for sequencing on the MGI DNBSEQ-G400RS instrument (MGI, Shenzhen, China). Total RNA was first treated with Globin-Zero Gold Kit (Illumina, San Diego, CA, USA) for rRNA depletion and globin mRNA reduction. The ds-cDNA library preparation is in line with the Illumina RNA-seq protocol described in the above section. The ds-cDNAs were then heat denatured and circularised by the splint oligo sequence to generate the single strand circle DNA followed by rolling circle replication to create DNA nanoballs (DNB) for processing on the MGI DNBSEQ-G400RS. The same 50 samples used for the GeneChip™ and Illumina RNA-seq profiling were again analysed on this platform. These samples were sequenced on 6 flowcells at 2x100 bp read length aimed at a sequencing depth of 64 M reads. Raw sequences were processed for quality assessment, alignment, transcripts quantification and DGE analysis as described above in the "**RNA-seq on Illumina NextSeq500**" section. The same cut-offs as in the Illumina RNA-seq section for defining a significant result were applied (i.e. a fold change of 1.2 bounded to the default *s*-value < 0.005).

Gene set enrichment analyses in GSEA and Reactome

The pathway enrichment analysis was performed in accordance with recommendations from *ref* [*37*], where appropriate. Specifically, normalised RNA-seq counts (outputted from DESeq2 “counts” function with the argument “normalized=TRUE”) and normalised microarray gene expression values were subjected to gene set enrichment analysis using GSEA (v4.0.3)[11, 12] by examining the Molecular Signatures Database (MSigDB)[11, 13] Hallmark (H; containing 50 gene sets)[14] and Gene Ontology (C5; BP: subset of GO biological processes containing 7,573 gene sets)[15, 16] collections of functional gene sets. Low count genes (by removing genes with counts below 5 in at least 4 samples) and genes with unidentified mappings from RNA-seq, and control probes, low-quality probes and probes with unidentified mappings from microarray analyses were excluded from the expression datasets prior to the GSEA. A standard GSEA run was applied for each dataset by performing 1,000 phenotype permutations and by collapsing the Ensembl IDs and probe IDs to gene symbols by mapping to their corresponding chip platforms available from the MSigDB database (i.e. Human_ENSEMBL_Gene_ID_MSigDB.v7.2.chip for RNA-seq, Human_AFFY_hta_2_0_MSigDB.v7.2.chip for GeneChip^{TM} HTA2.0 and Human_Illumina_HumanHT_12_v4_Array_MSigDB.v7.2.chip for Illumina BeadChip). Other main parameters used in a GSEA run included the default ranking metric “Signal2Noise”, gene set size filters (15-200 for H, and 10-500 for C5) and collapsing mode (“Sum_of_probes” for RNA-seq, and “Max_probe” for microarray). Default values were used for other fields of the GSEA run. EnrichmentMap App[38] was used for creating biological networks of the GSEA pathways (pathway FDR<0.1, nominal P<0.05 and Jaccard Overlap coefficient >0.375 with combined constant k=0.5) and AutoAnnotate App[39] for gene sets annotation and clustering (MCL Cluster annotation) in Cytoscape (v3.8.0)[40]. The most significantly enriched gene set was used to label a gene set cluster, characterised by the normalised enrichment score (NES). Raw counts from the RNA-seq (outputted from DESeq2 “counts” function by setting “normalized=FALSE”), and normalised and log2 transformed gene expression values from microarray analyses were uploaded onto Reactome (v73)[20] for quantitative pathway analysis (ReactomeGSA) using the PADOG algorithm[41, 42] for gene expression visualisation in pathway diagrams. Protein-protein interactors derived from the IntAct database[43] with the IntAct score ≥ 0.556 (of medium to high confidence interactions) were included in the analysis to improve the Reactome pathway coverage. For consistency, these expression datasets were collapsed to gene symbols using the “Collapse Dataset” tool in the GSEA software prior to the ReactomeGSA.

Cross-platform DGE comparison

Direct comparisons for the coding gene features identified across MGI DNBSEQ-G400RS, Illumina NextSeq 500 and GeneChip^{TM} HTA2.0 platforms in the 10 subjects (comprised of 50 samples) were carried out on the differentially expressed genes following the formal DESeq2/limma DGE analyses. A sankey diagram was plotted for visualisation of the DGE results using the “ggalluvial” package[44]. The cross-platform correlations were computed using the “ggscatter” function in the package “ggpubr”[45]. “ggplot2”[46] and “cowplot”[47] packages were used for creating publication-quality figures, where appropriate. Leading edge genes from the significantly expressed GSEA pathways (derived from MGI DNBSEQ-G400RS, GeneChip^{TM} HTA2.0 and HumanHT-12 v4.0 Expression BeadChip; including all pathways with the NES>1.9 or the representative pathway of a gene set cluster when the NES<1.9) were extracted and compared to the DGE genes to generate common sets of genes identified by both the GSEA and DGE analyses. These genes were then overlapped with the DGE results obtained from the

Illumina NextSeq 500 platform for confirmation. The interaction networks among pathway genes were defined by expression overlay with the Reactome pathway diagrams, focusing on the networks represented by the confirmed genes above. The final lists of genes were obtained by extracting all significantly altered genes and their interactors involved in these Reactome networks and by matching them back to the differentially expressed genes resulted from the formal DESeq2/limma analyses.

Data availability: Note all data will be made available and deposited into appropriate repositories at publication, including raw RNA-seq data, raw microarray data, and code required to reproduce all analyses. Full data access may be provided to reviewers on request during manuscript reviewing.

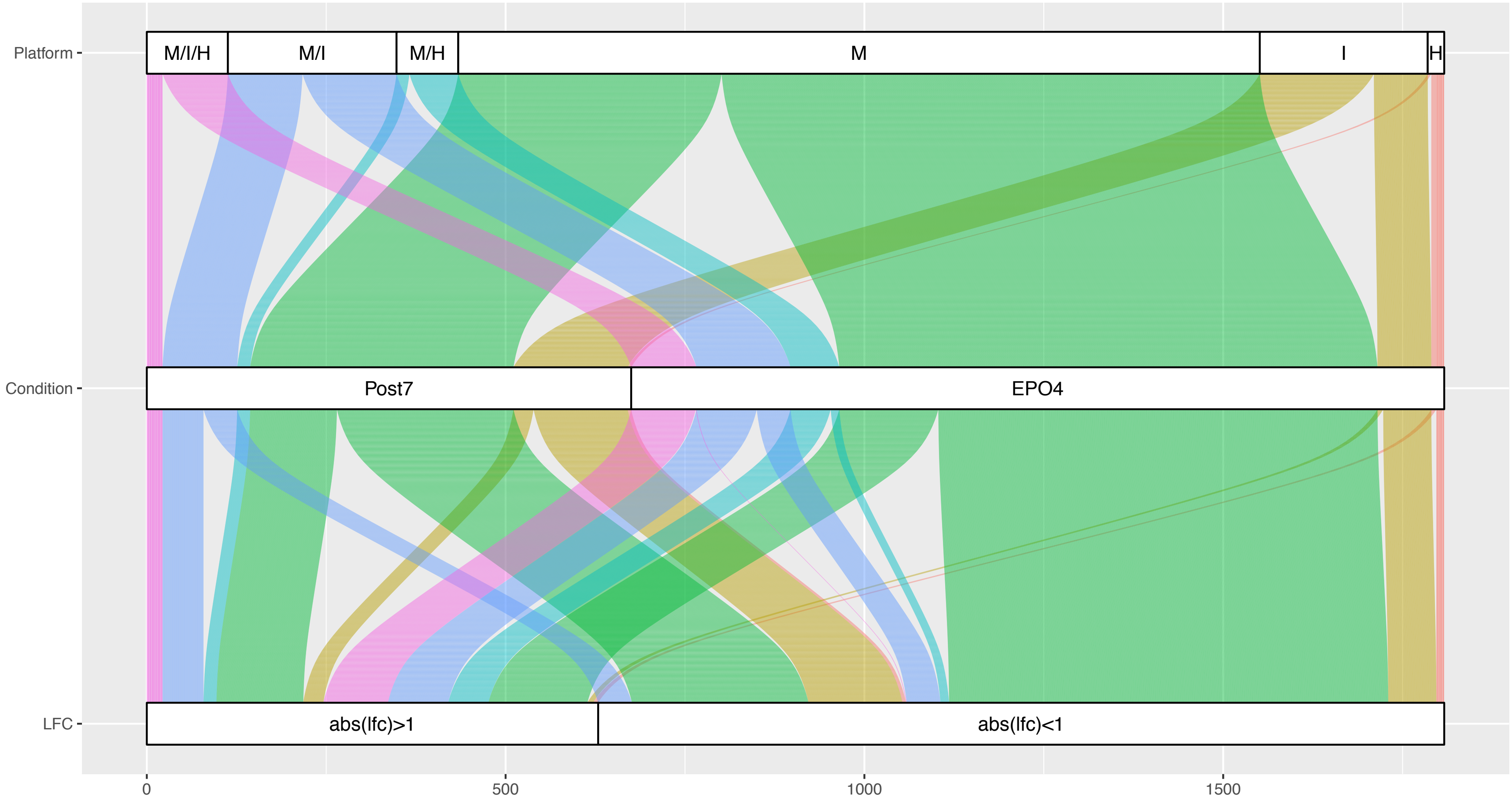


Fig. 1. Sankey diagram showing the flow of the differentially expressed gene features stratified by platform, biological condition and absolute log$_2$-transformed fold changes. M/I/H: MGI RNA-seq/Illumina RNA-seq/HTA2.0; M/I: MGI RNA-seq/Illumina RNA-seq; M/H: MGI RNA-seq/HTA2.0; M: MGI RNA-seq; I: Illumina RNA-seq; and H: HTA2.0. abs(lfc): absolute log$_2$-transformed fold change. The colour coded band represents a detection platform or a combination of the detection platforms. The

wider the band, the higher number of the identified features on a platform or across platforms. The x-axis represents the number of identified features captured on each platform. Note, for M/I/H, M/I, and M/H, that biological magnitude of the features used for stratification is based on the MGI RNA-seq DGE results. Thirty-four identified non-protein coding transcript clusters on the GeneChip are removed for the purposes of cross-platform comparison.

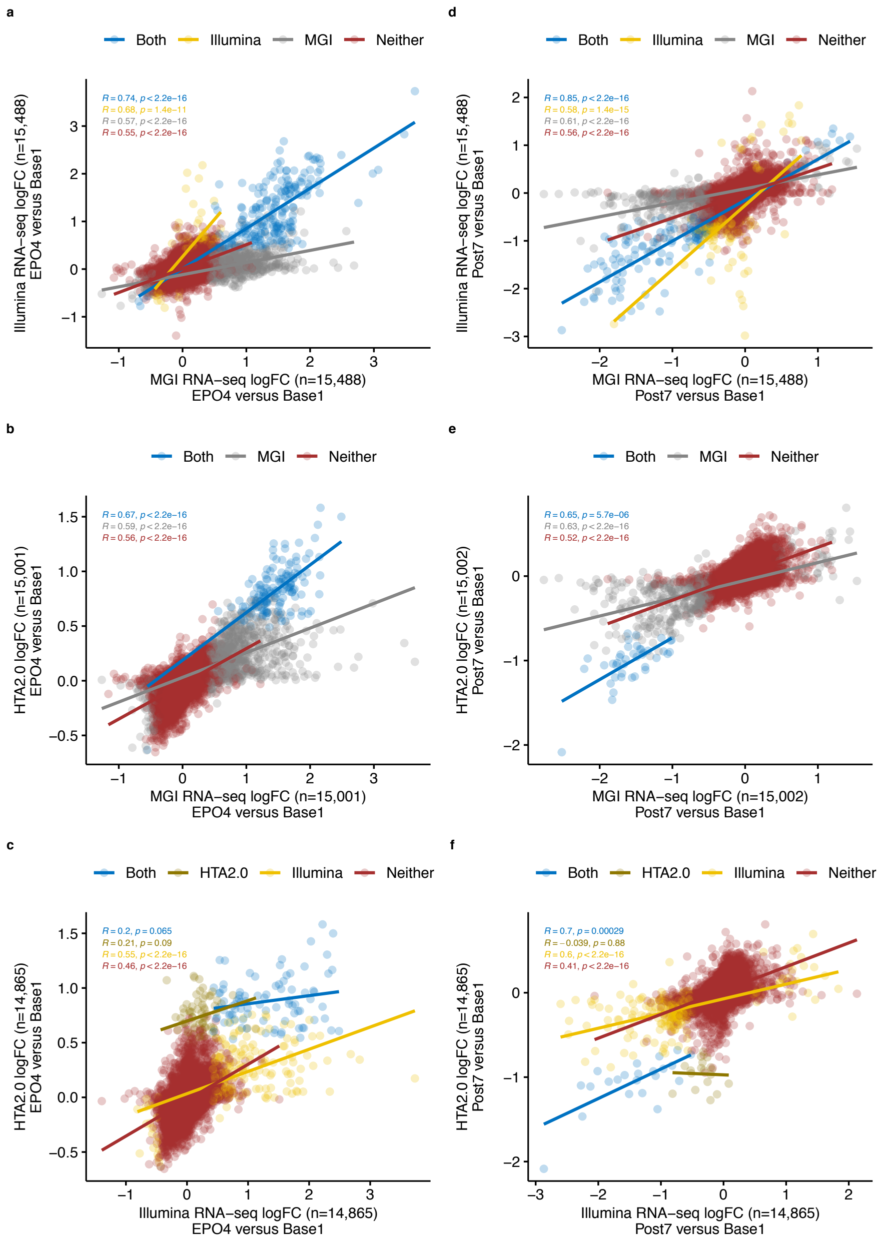
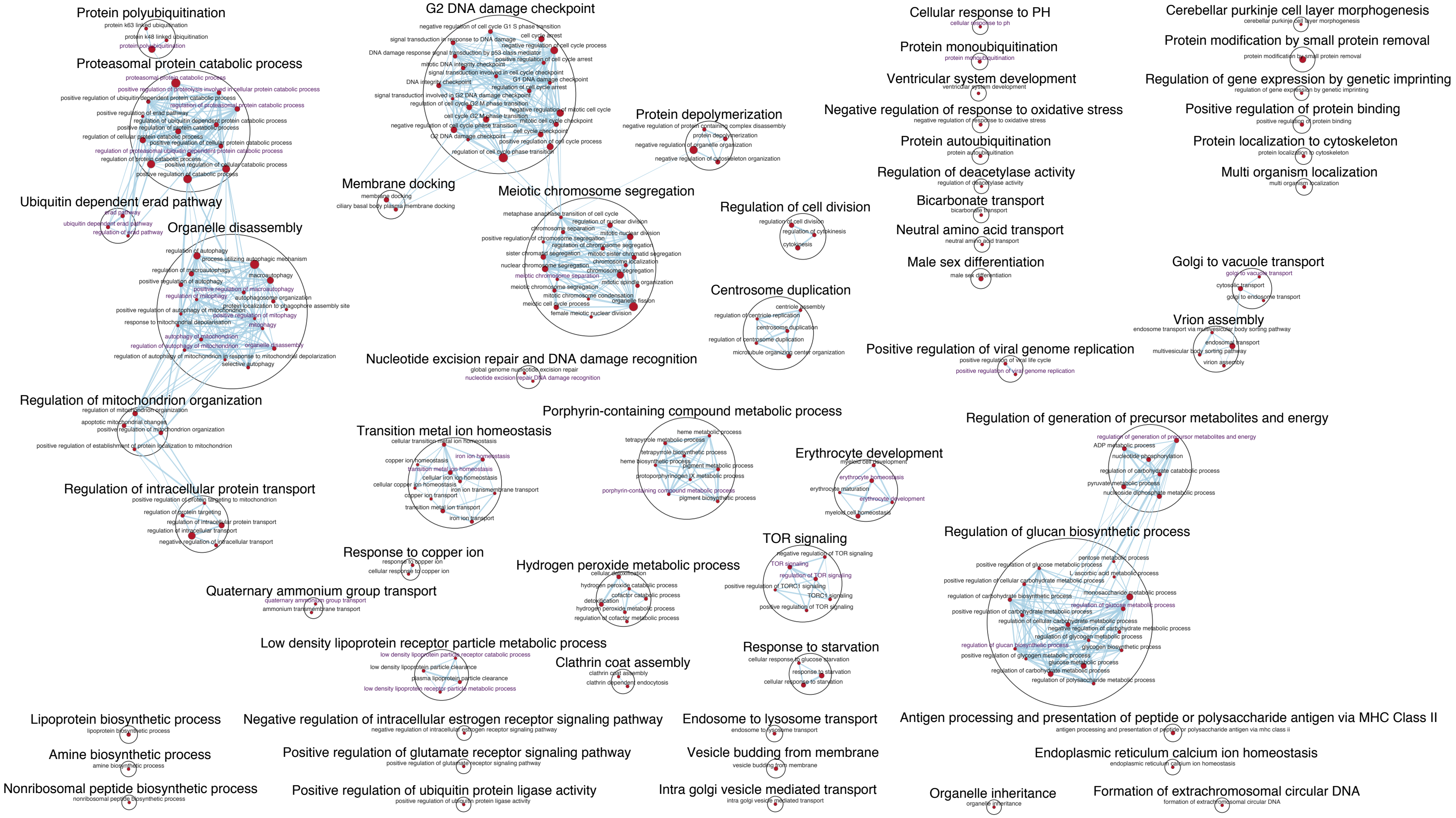


Fig. 2. Cross-platform gene expression correlation analyses of log$_2$-transformed fold changes of all identified gene features. a-c Genes identified when compared the level of expression between EPO4 and Base1 among the platform pairs in Illumina-MGI RNA-seq (**a**), GeneChip^{TM} HTA2.0-MGI RNA-seq (**b**), GeneChip^{TM} HTA2.0-Illlumina RNA-seq (**c**). **d-f** Genes identified when compared the level of expression between Post7 and Base1 among the

platform pairs in Illumina-MGI RNA-seq (**d**), GeneChip^{TM} HTA2.0-MGI RNA-seq (**e**), GeneChip^{TM} HTA2.0-Illlumina RNA-seq (**f**). Genes identified as differentially expressed by each pair are plotted in blue; genes that are only differentially expressed in Illumina RNA-seq, MGI RNA-seq or GeneChip^{TM} HTA2.0 are plotted in yellow, grey and dijon, respectively; genes not identified as differentially expressed by a pair are plotted in red. For simplicity, the maximum expression value of a gene was used when multiple mapping of transcripts to the same gene occurred. *FOXO3B* is only differentially expressed in GeneChip^{TM} HTA2.0 when compared to the MGI RNA-seq findings in (**b**), thus it has been removed from the correlation analysis. R: Pearson's *r*. LogFC: $\log_2$-transformed fold change.



Protein polyubiquitination
protein k63 linked ubiquitination
protein k48 linked ubiquitination
protein polyubiquitination
Proteasomal protein catabolic process
proteasomal protein catabolic process
positive regulation of proteolysis involved in cellular protein catabolic process
positive regulation of ubiquitin dependent protein catabolic process
regulation of proteasomal protein catabolic process
positive regulation of erad pathway
regulation of ubiquitin dependent protein catabolic process
positive regulation of protein catabolic process
regulation of cellular protein catabolic process
positive regulation of cellular protein catabolic process
regulation of proteasomal ubiquitin dependent protein catabolic process
regulation of protein catabolic process
positive regulation of cellular catabolic process
positive regulation of catabolic process
Ubiquitin dependent erad pathway
erad pathway
ubiquitin dependent erad pathway
regulation of erad pathway
Organelle disassembly
regulation of autophagy
process utilizing autophagic mechanism
regulation of macroautophagy
macroautophagy
positive regulation of autophagy
positive regulation of macroautophagy
regulation of mitophagy
autophagosome organization
protein localization to phagophore assembly site
positive regulation of autophagy of mitochondrion
positive regulation of mitophagy
response to mitochondrial depolarisation
mitophagy
autophagy of mitochondrion
regulation of autophagy of mitochondrion
organelle disassembly
regulation of autophagy of mitochondrion in response to mitochondrial depolarization
selective autophagy
Regulation of mitochondrion organization
regulation of mitochondrion organization
apoptotic mitochondrial changes
positive regulation of mitochondrion organization
positive regulation of establishment of protein localization to mitochondrion
Regulation of intracellular protein transport
positive regulation of protein targeting to mitochondrion
regulation of protein targeting
regulation of intracellular protein transport
regulation of intracellular transport
negative regulation of intracellular transport
G2 DNA damage checkpoint
negative regulation of cell cycle G1 S phase transition
signal transduction in response to DNA damage
cell cycle arrest
negative regulation of cell cycle process
DNA damage response signal transduction by p53 class mediator
positive regulation of cell cycle arrest
mitotic DNA integrity checkpoint
signal transduction involved in cell cycle checkpoint
DNA integrity checkpoint
G1 DNA damage checkpoint
regulation of cell cycle arrest
signal transduction involved in G2 DNA damage checkpoint
regulation of cell cycle G2 M phase transition
negative regulation of mitotic cell cycle
cell cycle G2 M phase transition
mitotic cell cycle checkpoint
negative regulation of cell cycle phase transition
cell cycle checkpoint
G2 DNA damage checkpoint
positive regulation of cell cycle process
regulation of cell cycle phase transition
Membrane docking
membrane docking
ciliary basal body plasma membrane docking
Protein depolymerization
negative regulation of protein containing complex disassembly
protein depolymerization
negative regulation of organelle organization
negative regulation of cytoskeleton organization
Meiotic chromosome segregation
metaphase anaphase transition of cell cycle
regulation of nuclear division
chromosome separation
mitotic nuclear division
positive regulation of chromosome segregation
regulation of chromosome segregation
sister chromatid segregation
mitotic sister chromatid segregation
nuclear chromosome segregation
chromosome localization
chromosome segregation
meiotic chromosome separation
mitotic spindle organization
meiotic chromosome segregation
mitotic chromosome condensation
organelle fission
meiotic cell cycle process
female meiotic nuclear division
Nucleotide excision repair and DNA damage recognition
global genome nucleotide excision repair
nucleotide excision repair DNA damage recognition
Regulation of cell division
regulation of cell division
regulation of cytokinesis
cytokinesis
Centrosome duplication
centriole assembly
regulation of centriole replication
centrosome duplication
regulation of centrosome duplication
microtubule organizing center organization
Cellular response to PH
cellular response to ph
Protein monoubiquitination
protein monoubiquitination
Ventricular system development
ventricular system development
Negative regulation of response to oxidative stress
negative regulation of response to oxidative stress
Protein autoubiquitination
protein autoubiquitination
Regulation of deacetylase activity
regulation of deacetylase activity
Bicarbonate transport
bicarbonate transport
Neutral amino acid transport
neutral amino acid transport
Male sex differentiation
male sex differentiation
Positive regulation of viral genome replication
positive regulation of viral life cycle
positive regulation of viral genome replication
Cerebellar purkinje cell layer morphogenesis
cerebellar purkinje cell layer morphogenesis
Protein modification by small protein removal
protein modification by small protein removal
Regulation of gene expression by genetic imprinting
regulation of gene expression by genetic imprinting
Positive regulation of protein binding
positive regulation of protein binding
Protein localization to cytoskeleton
protein localization to cytoskeleton
Multi organism localization
multi organism localization
Golgi to vacuole transport
golgi to vacuole transport
cytosolic transport
golgi to endosome transport
Vrion assembly
endosome transport via multivesicular body sorting pathway
endosomal transport
multivesicular body sorting pathway
virion assembly
Transition metal ion homeostasis
cellular transition metal ion homeostasis
copper ion homeostasis
iron ion homeostasis
transition metal ion homeostasis
cellular iron ion homeostasis
cellular copper ion homeostasis
iron ion transmembrane transport
copper ion transport
transition metal ion transport
iron ion transport
Porphyrin-containing compound metabolic process
heme metabolic process
tetrapyrrole metabolic process
tetrapyrrole biosynthetic process
heme biosynthetic process
pigment metabolic process
protoporphyrinogen IX metabolic process
porphyrin-containing compound metabolic process
pigment biosynthetic process
Erythrocyte development
myeloid cell development
erythrocyte homeostasis
erythrocyte maturation
erythrocyte development
myeloid cell homeostasis
Regulation of generation of precursor metabolites and energy
regulation of generation of precursor metabolites and energy
ADP metabolic process
nucleotide phosphorylation
regulation of carbohydrate catabolic process
pyruvate metabolic process
nucleoside diphosphate metabolic process
Response to copper ion
response to copper ion
cellular response to copper ion
Quaternary ammonium group transport
quaternary ammonium group transport
ammonium transmembrane transport
Hydrogen peroxide metabolic process
cellular detoxification
hydrogen peroxide catabolic process
cofactor catabolic process
detoxification
hydrogen peroxide metabolic process
regulation of cofactor metabolic process
TOR signaling
negative regulation of TOR signaling
TOR signaling
regulation of TOR signaling
positive regulation of TORC1 signaling
TORC1 signaling
positive regulation of TOR signaling
Regulation of glucan biosynthetic process
pentose metabolic process
positive regulation of glucose metabolic process
L ascorbic acid metabolic process
positive regulation of cellular carbohydrate metabolic process
regulation of carbohydrate biosynthetic process
monosaccharide metabolic process
regulation of glucose metabolic process
positive regulation of carbohydrate metabolic process
regulation of cellular carbohydrate metabolic process
negative regulation of carbohydrate metabolic process
regulation of glycogen metabolic process
regulation of glucan biosynthetic process
glycogen biosynthetic process
positive regulation of glycogen metabolic process
glucose metabolic process
regulation of carbohydrate metabolic process
regulation of polysaccharide metabolic process
Low density lipoprotein receptor particle metabolic process
low density lipoprotein particle receptor catabolic process
low density lipoprotein particle clearance
plasma lipoprotein particle clearance
low density lipoprotein receptor particle metabolic process
Clathrin coat assembly
clathrin coat assembly
clathrin dependent endocytosis
Response to starvation
cellular response to glucose starvation
response to starvation
cellular response to starvation
Lipoprotein biosynthetic process
lipoprotein biosynthetic process
Amine biosynthetic process
amine biosynthetic process
Nonribosomal peptide biosynthetic process
nonribosomal peptide biosynthetic process
Negative regulation of intracellular estrogen receptor signaling pathway
negative regulation of intracellular estrogen receptor signaling pathway
Positive regulation of glutamate receptor signaling pathway
positive regulation of glutamate receptor signaling pathway
Positive regulation of ubiquitin protein ligase activity
positive regulation of ubiquitin protein ligase activity
Endosome to lysosome transport
endosome to lysosome transport
Vesicle budding from membrane
vesicle budding from membrane
Intra golgi vesicle mediated transport
intra golgi vesicle mediated transport
Antigen processing and presentation of peptide or polysaccharide antigen via MHC Class II
antigen processing and presentation of peptide or polysaccharide antigen via mhc class ii
Endoplasmic reticulum calcium ion homeostasis
endoplasmic reticulum calcium ion homeostasis
Organelle inheritance
organelle inheritance
Formation of extrachromosomal circular DNA
formation of extrachromosomal circular DNA

Fig. 3. Biological network of the MGI RNA-seq dataset following Gene Ontology (biological process) gene set enrichment analysis in GSEA (v4.0.3) and visualisation in Cytoscape (3.8.0). Each circle (node) represents a gene set and two nodes are

connected by lines (edges) indicating shared genes. The size of a node and width of an edge are proportional to the number of genes enriched in a gene set and the number of genes shared between gene sets, respectively. Gene sets that are similar were annotated and clustered to form a biological theme using the AutoAnnotate App in Cytoscape. The most significantly enriched gene set is used to label a gene set cluster, defined by NES. Red node: gene set enriched in EPO4. Purple node label: top gene sets with NES > 1.90. The enrichment map was created with pathway FDR < 0.1, nominal *P* < 0.05 and Jaccard Overlap coefficient > 0.375 with combined constant k = 0.5.

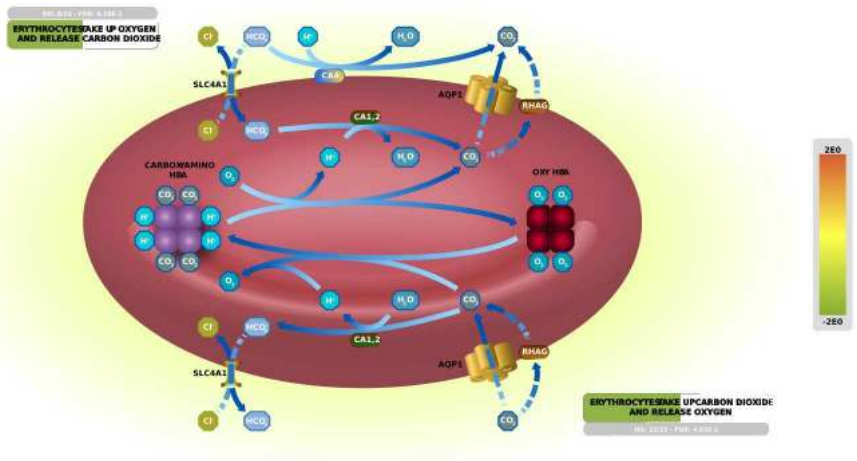

ERYTHROCYTES TAKE UP OXYGEN AND RELEASE CARBON DIOXIDE
Cl⁻
HCO₃⁻
H⁺
H₂O
CO₂
CA4
SLC4A1
AQP1
RHAG
CA1,2
CARBOXYAMINO HBA
O₂
OXY HBA
ERYTHROCYTES TAKE UP CARBON DIOXIDE AND RELEASE OXYGEN
2E0
-2E0

Fig. 4. Enhanced high-level Reactome pathway diagram for O_2/CO_2 exchange in erythrocytes[48] by expression overlay with the MGI RNA-seq Post7 dataset. This high-level diagram represents two subpathways, namely erythrocyte take up oxygen and release carbon dioxide and erythrocyte take up carbon dioxide and release oxygen. The green band indicates the proportion of the pathway that is represented in the MGI RNA-seq Post7 dataset, and the colour (green) represents the down-regulation of the pathway genes. The grey bar contains the information for the number of pathway entities in the query dataset, the total number of the pathway entities, and the FDR corrected over-representation probability.

References:

1. Shendure, J. et al., DNA sequencing at 40: past, present and future. *Nature*. **550**, 345-353 (2017).

2. Manolio, T. A. et al., Genomic Medicine Year in Review: 2019. *Am. J. Hum. Genet.* **105**, 1072-1075 (2019).

3. Durussel, J. et al., Haemoglobin mass and running time trial performance after recombinant human erythropoietin administration in trained men. *PLoS One*. **8**, e56151 (2013).

4. Durussel, J. et al., Blood transcriptional signature of recombinant human erythropoietin administration and implications for antidoping strategies. *Physiol. Genomics*. **48**, 202-209 (2016).

5. Kim, D., Paggi, J. M., Park, C., Bennett, C., Salzberg, S. L., Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907-915 (2019).

6. Schneider, V. A. et al., Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849-864 (2017).

7. Wang, L., Wang, S., Li, W., RSeQC: quality control of RNA-seq experiments. *Bioinformatics*. **28**, 2184-2185 (2012).

8. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., Kingsford, C., Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*. **14**, 417-419 (2017).

9. Love, M. I., Huber, W., Anders, S., Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

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

11. Subramanian, A. et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545-15550 (2005).

12. Mootha, V. K. et al., PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267-273 (2003).

13. Liberzon, A. et al., Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. **27**, 1739-1740 (2011).

14. Liberzon, A. et al., The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*. **1**, 417-425 (2015).

15. Ashburner, M. et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25-29 (2000).

16. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330-d338 (2019).

17. Ghezzi, P., Brines, M., Erythropoietin as an antiapoptotic, tissue-protective cytokine. *Cell Death Differ.* **11 Suppl 1**, S37-44 (2004).

18. Jelkmann, W., Regulation of erythropoietin production. *J. Physiol.* **589**, 1251-1258 (2011).

19. Maiese, K., Erythropoietin and diabetes mellitus. *World J. Diabetes*. **6**, 1259-1273 (2015).

20. Jassal, B. et al., The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498-d503 (2020).

21. Shi, W., Oshlack, A., Smyth, G. K., Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. *Nucleic Acids Res.* **38**, e204 (2010).

22. Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S., Smyth, G. K., Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Ann. Appl. Stat*. **10**, 946-963 (2016).

23. Dunning, M., Lynch, A., Eldridge, M., illuminaHumanv4.db: Illumina HumanHT12v4 annotation data (chip illuminaHumanv4). R package version 1.26.0. (2015).

24. Vaes, E., Khan, M., Mombaerts, P., Statistical analysis of differential gene expression relative to a fold change threshold on NanoString data of mouse odorant receptor genes. *BMC Bioinformatics*. **15**, 39 (2014).

25. Klaus, B., Reisenauer, S., An end to end workflow for differential gene expression using Affymetrix microarrays. *F1000Res*. **5**, 1384 (2016).

26. Carvalho, B. S., Irizarry, R. A., A framework for oligonucleotide microarray preprocessing. *Bioinformatics*. **26**, 2363-2367 (2010).

27. MacDonald, J. W., hta20transcriptcluster.db: Affymetrix hta20 annotation data (chip hta20transcriptcluster). R package version 8.7.0. (2017).

28. Andrews, S., FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (2010).

29. Wingett, S. W., Andrews, S., FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res*. **7**, 1338 (2018).

30. Ensembl Archive Release 94 (October 2018). Available at: http://oct2018.archive.ensembl.org/index.html. (2018).

31. Soneson, C., Love, M. I., Robinson, M. D., Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res*. **4**, 1521 (2015).

32. Carlson, M., org.Hs.eg.db: Genome wide annotation for Human. R package version 3.8.2., (2019).

33. Ewels, P., Magnusson, M., Lundin, S., Käller, M., MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. **32**, 3047-3048 (2016).

34. Love, M. I., Anders, S., Huber, W., Analyzing RNA-seq data with DESeq2. Available at: https://www.bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html - references. (2020).

35. Leek, J. T. et al., sva: Surrogate Variable Analysis. R package version 3.38.0. (2020).

36. Zhu, A., Ibrahim, J. G., Love, M. I., Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*. **35**, 2084-2092 (2019).

37. Reimand, J. et al., Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* **14**, 482-517 (2019).

38. Merico, D., Isserlin, R., Stueker, O., Emili, A., Bader, G. D., Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One*. **5**, e13984 (2010).

39. Kucera, M., Isserlin, R., Arkhangorodsky, A., Bader, G. D., AutoAnnotate: A Cytoscape app for summarizing networks with semantic annotations. *F1000Res*. **5**, 1717 (2016).

40. Shannon, P. et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. **13**, 2498-2504 (2003).

41. Tarca, A. L., Draghici, S., Bhatti, G., Romero, R., Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*. **13**, 136 (2012).

42. Griss, J. et al., ReactomeGSA - Efficient Multi-Omics Comparative Pathway Analysis. *Mol. Cell. Proteomics*. (2020).

43. Orchard, S. et al., The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358-363 (2014).

44. Brunson, J. C., ggalluvial: Layered Grammar for Alluvial Plots. *J. Open Source Softw*. **5**, 49 (2017).

45. Kassambara, A., ggpubr: 'ggplot2' based publication ready plots. R package version 0.4.0. (2020).

46. Wickham, H., Elegant graphics for data analysis. Springer-Verlag New York. (2016).

47. Wilke, C. O., cowplot: streamlined plot theme and plot annotations for 'ggplot2'. R package version 1.1.1. (2020).

48. O2/CO2 exchange in erythrocytes. Reactome, released 2012-06-12, doi:10.3180/REACT_120969.1 (11/11/20).

Acknowledgments:

We thank BGI Hong Kong for conducting the DNB-seq. This work was funded by the World Anti-Doping Agency (WADA) grant no. 08C19YP and ISF15E10YP to YPP. GIA was supported by a fellowship from the Office of Academic Affiliations at the United States Veterans Health Administration.

Author contributions:

Conceptualization: GW, YPP; Formal analysis: GW; Funding acquisition: YPP; Investigation: GW, TK; Methodology: GW, YPP; Project administration: AC, QM; Resources: GIA, JL, MBG; Supervision: YPP; Validation: GW; Visualization: GW; Writing – original draft: GW; Writing – review & editing: AH, FMG, GIA, JL, MBG, YPP, GW.

Competing interests:

The authors declare no competing interests.

Materials & Correspondence:

Correspondence and material requests should be addressed to GW and YPP.

Figure legends:

Fig. 1. Sankey diagram [44] showing the flow of the differentially expressed gene features stratified by platform, biological condition and absolute $\log_2$-transformed fold changes. M/I/H: MGI RNA-seq/Illumina RNA-seq/HTA2.0; M/I: MGI RNA-seq/Illumina RNA-seq; M/H: MGI RNA-seq/HTA2.0; M: MGI RNA-seq; I: Illumina RNA-seq; and H: HTA2.0. abs(lfc): absolute $\log_2$-transformed fold change. The colour coded band represents a detection platform or a combination of the detection platforms. The wider the band, the higher number of the identified features on a platform or across platforms. The x-axis represents the number of identified features captured on each platform. Note, for M/I/H, M/I, and M/H, that biological magnitude of the features used for stratification is based on the MGI RNA-seq DGE results. Thirty-four identified non-protein coding transcript clusters on the GeneChip are removed for the purposes of cross-platform comparison.

Fig. 2. Cross-platform gene expression correlation analyses of $\log_2$-transformed fold changes of all identified gene features. a-c Genes identified when compared the level of expression between EPO4 and Base1 among the platform pairs in Illumina-MGI RNA-seq (**a**), GeneChip^{TM} HTA2.0-MGI RNA-seq (**b**), GeneChip^{TM} HTA2.0-Illlumina RNA-seq (**c**). **d-f** Genes identified when compared the level of expression between Post7 and Base1 among the platform pairs in Illumina-MGI RNA-seq (**d**), GeneChip^{TM} HTA2.0-MGI RNA-seq (**e**), GeneChip^{TM} HTA2.0-Illlumina RNA-seq (**f**). Genes identified as differentially expressed by each pair are plotted in blue; genes that are only differentially expressed in Illumina RNA-seq, MGI RNA-seq or GeneChip^{TM} HTA2.0 are plotted in yellow, grey and dijon, respectively; genes not identified as differentially expressed by a pair are plotted in red. For simplicity, the maximum expression value of a gene was used when multiple mapping of transcripts to the same

gene occurred. FOXO3B is only differentially expressed in GeneChip^{TM} HTA2.0 when compared to the MGI RNA-seq findings in (**b**), thus it has been removed from the correlation analysis. R: Pearson's r. LogFC: $\log_2$-transformed fold change.

Fig. 3. Biological network of the MGI RNA-seq dataset following Gene Ontology (biological process) gene set enrichment analysis in GSEA (v4.0.3) and visualisation in Cytoscape (3.8.0) [40]. Each circle (node) represents a gene set and two nodes are connected by lines (edges) indicating shared genes. The size of a node and width of an edge are proportional to the number of genes enriched in a gene set and the number of genes shared between gene sets, respectively. Gene sets that are similar were annotated and clustered to form a biological theme using the AutoAnnotate App [39] in Cytoscape. The most significantly enriched gene set is used to label a gene set cluster, defined by NES. Red node: gene set enriched in EPO4. Purple node label: top gene sets with NES > 1.90. The enrichment map was created with pathway FDR < 0.1, nominal $P < 0.05$ and Jaccard Overlap coefficient > 0.375 with combined constant k = 0.5.

Fig. 4. Enhanced high-level Reactome pathway diagram for O_2/CO_2 exchange in erythrocytes [48] by expression overlay with the MGI RNA-seq Post7 dataset. This high-level diagram represents two subpathways, namely erythrocyte take up oxygen and release carbon dioxide and erythrocyte take up carbon dioxide and release oxygen. The green band indicates the proportion of the pathway that is represented in the MGI RNA-seq Post7 dataset, and the colour (green) represents the down-regulation of the pathway genes. The grey bar contains the information for the number of pathway entities in the query dataset, the total number of the pathway entities, and the FDR corrected over-representation probability.

Table 1. Summary of the number of transcriptomic features available for the DGE analysis across the four gene expression detection platforms.

	MGI DNBSEQ-G400RS	Illumina NextSeq 500	GeneChip™ HTA2.0	Illumina HumanHT-12 v4 Expression BeadChip
Number of samples	50	48	49	143
Number of transcriptomic features following RNA-seq quantification (Salmon) or on the array	175,775	175,775	285,543	47,286
Number of identified features available for the DGE analysis	16,738[g]	16,581[g]	29,517[tc]	10,622[t]

DGE: differential gene expression. g: protein-coding gene features. tc: protein coding and non-protein coding transcript clusters, loosely corresponding to genes. t: coding and non-coding transcripts. Two, one and one samples were removed from the DGE analyses due to human processing errors, sample quality and sampling issue in the Illumina RNA-seq, GeneChip and BeadChip datasets, respectively.

Table 2. Transcript annotation and filtering of the RNA-seq and microarray data prior to the DGE analysis.

	MGI DNBSEQ-G400RS	Illumina NextSeq 500	GeneChip™ HTA2.0	Illumina HumanHT-12 v4 Expression BeadChip
Annotation database (N=the number of transcriptomic features)	Org.Hs.eg.db (N=175,775 transcripts following Salmon transcription quantification, aggregated into 37,788 genes using Ensembl 94 annotation)	Org.Hs.eg.db (N=175,775 transcripts following Salmon transcription quantification, aggregated into 37,788 genes using Ensembl 94 annotation)	hta20transcriptcluster.db (N=285,543 transcripts, corresponding to 67,480 protein-coding and non-protein coding transcript clusters)	illuminaHumanv4.db (N=47,286 coding and non-coding transcripts)
Undetected probes	-	-	-	18,494
Low quality probes	-	-	-	6,900
Low-expressed genes (RNA-seq)	17,198	18,347	-	-
“NA” mapping to stable gene symbols	3,675	2,668	36,709	2,406
Multiple mapping to stable gene symbols	177	192	1,254	698
Low-expressed probes (microarray)[1]	-	-	0	8,166
Identified features available for DGE analysis	16,738[g]	16,581[g]	29,517[tc]	10,622[t]

NA: features with no gene symbols returned after annotation. DGE: differential gene expression. -: not applicable. [1]: low expressed probes were further removed following assessing the average log expression and the mean-variance relationship after fitting the linear model in limma microarray analysis. g: protein-coding gene features. tc: protein-coding and non-protein coding transcript clusters, loosely corresponding to genes. t: coding and non-coding transcript features.

Table 3. Summary of the number of significantly expressed transcriptomic features across all platforms.

	DGE thresholds	Base2 vs Base1	EPO3 vs Base1	EPO4 vs Base1	Post7 vs Base1	Up/down regulation
MGI DNBSEQ-G400RS (N = 50)	abs FC > 1.2 & *s*-value < 0.005	0	1	959	60	Up
		0	0	81	451	Down
Illumina NextSeq 500 (N = 48)	abs FC > 1.2 & *s*-value < 0.005	0	0	277	27	Up
		0	0	20	258	Down
GeneChip$^{\text{TM}}$ HTA2.0 (N = 49)	abs FC > 1.2 & FDR < 0.05	0	0	200	0	Up
		0	0	1	51	Down
Illumina HumanHT-12v4.0 Expression Beadchip (N = 143)	abs FC > 1.2 & FDR < 0.05	0	13	796	7	Up
		0	0	1,315	254	Down

DGE: differential gene expression. abs FC: absolute fold change. FDR: false discovery rate. The number of protein-coding gene features, and coding and non-coding transcript clusters and transcripts are reported following the RNA-seq, GeneChip and Beadchip DGE analyses, respectively.

Figures

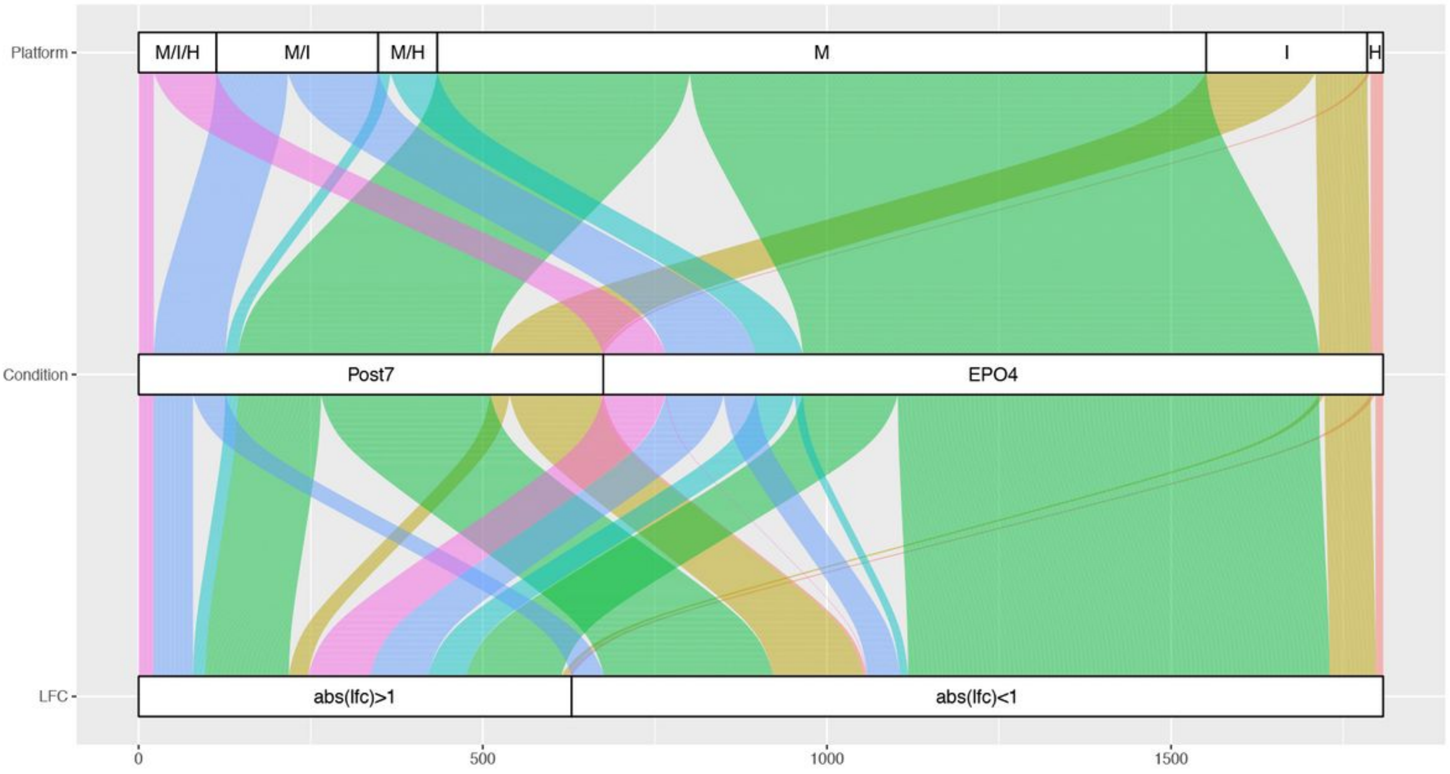


Figure 1

Sankey diagram 44 showing the flow of the differentially expressed gene features stratified by platform, biological condition and absolute log2-transformed fold changes. M/I/H: MGI RNA-seq/Illumina RNA-seq/HTA2.0; M/I: MGI RNA-seq/Illumina RNA-seq; M/H: MGI RNA-seq/HTA2.0; M: MGI RNA-seq; I: Illumina RNA-seq; and H: HTA2.0. abs(lfc): absolute log2-transformed fold change. The colour coded band represents a detection platform or a combination of the detection platforms. The wider the band, the higher number of the identified features on a platform or across platforms. The x-axis represents the number of identified features captured on each platform. Note, for M/I/H, M/I, and M/H, that biological magnitude of the features used for stratification is based on the MGI RNA-seq DGE results. Thirty-four identified non-protein coding transcript clusters on the GeneChip are removed for the purposes of cross-platform comparison.

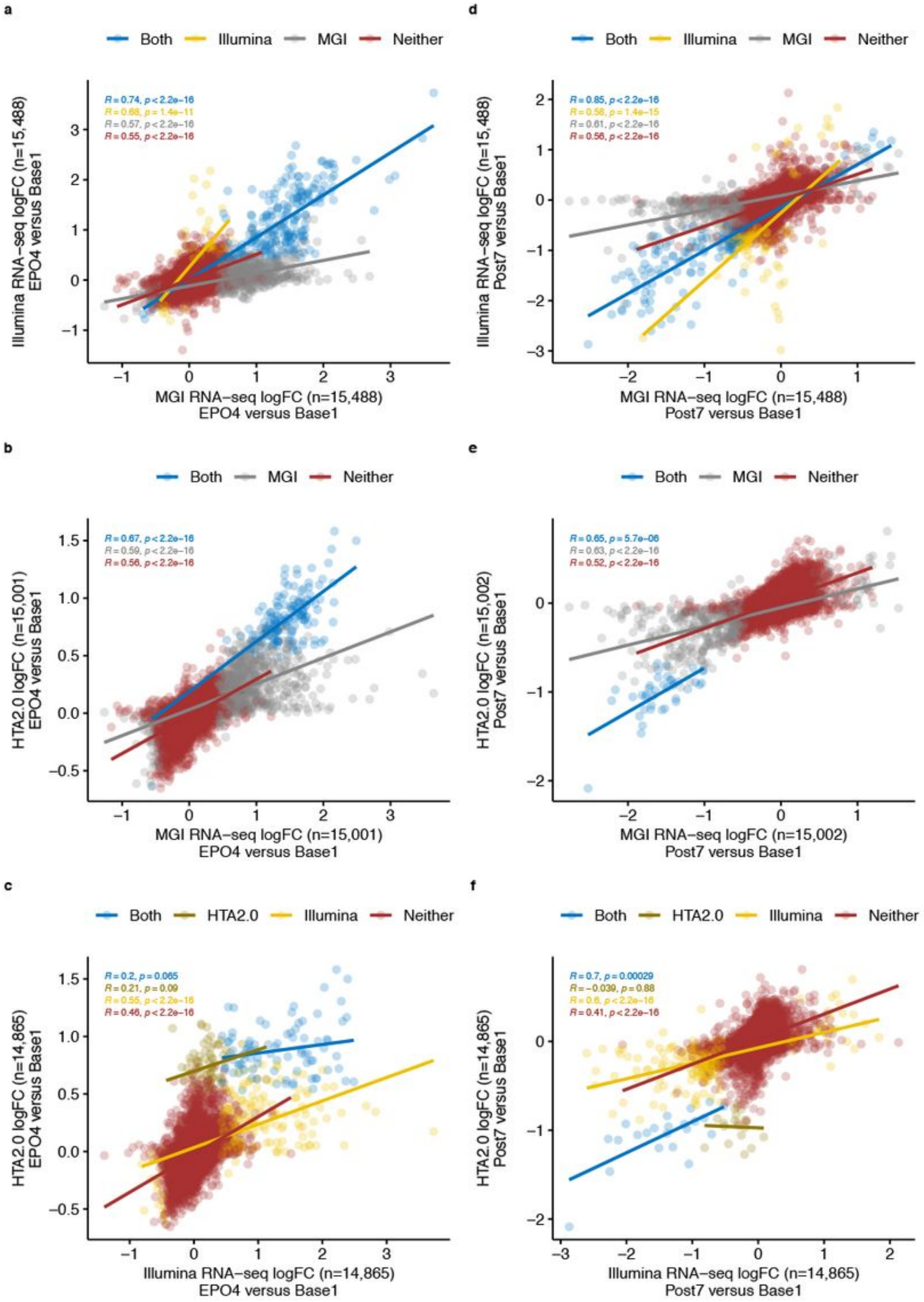


Figure 2

Cross-platform gene expression correlation analyses of log2-transformed fold changes of all identified gene features. a-c Genes identified when compared the level of expression between EPO4 and Base1 among the platform pairs in Illumina-MGI RNA-seq (a), GeneChipTM HTA2.0-MGI RNA-seq (b), GeneChipTM HTA2.0-Illlumina RNA-seq (c). d-f Genes identified when compared the level of expression between Post7 and Base1 among the platform pairs in Illumina-MGI RNA-seq (d), GeneChipTM HTA2.0-

MGI RNA-seq (e), GeneChipTM HTA2.0-IIllumina RNA-seq (f). Genes identified as differentially expressed by each pair are plotted in blue; genes that are only differentially expressed in Illumina RNA-seq, MGI RNA-seq or GeneChipTM HTA2.0 are plotted in yellow, grey and dijon, respectively; genes not identified as differentially expressed by a pair are plotted in red. For simplicity, the maximum expression value of a gene was used when multiple mapping of transcripts to the same gene occurred. FOXO3B is only differentially e xpressed in GeneChipTM HTA2.0 when compared to the MGI RNA-seq findings in (b), thus it has been removed from the correlation analysis. R: Pearson's r. LogFC: log2-transformed fold change.

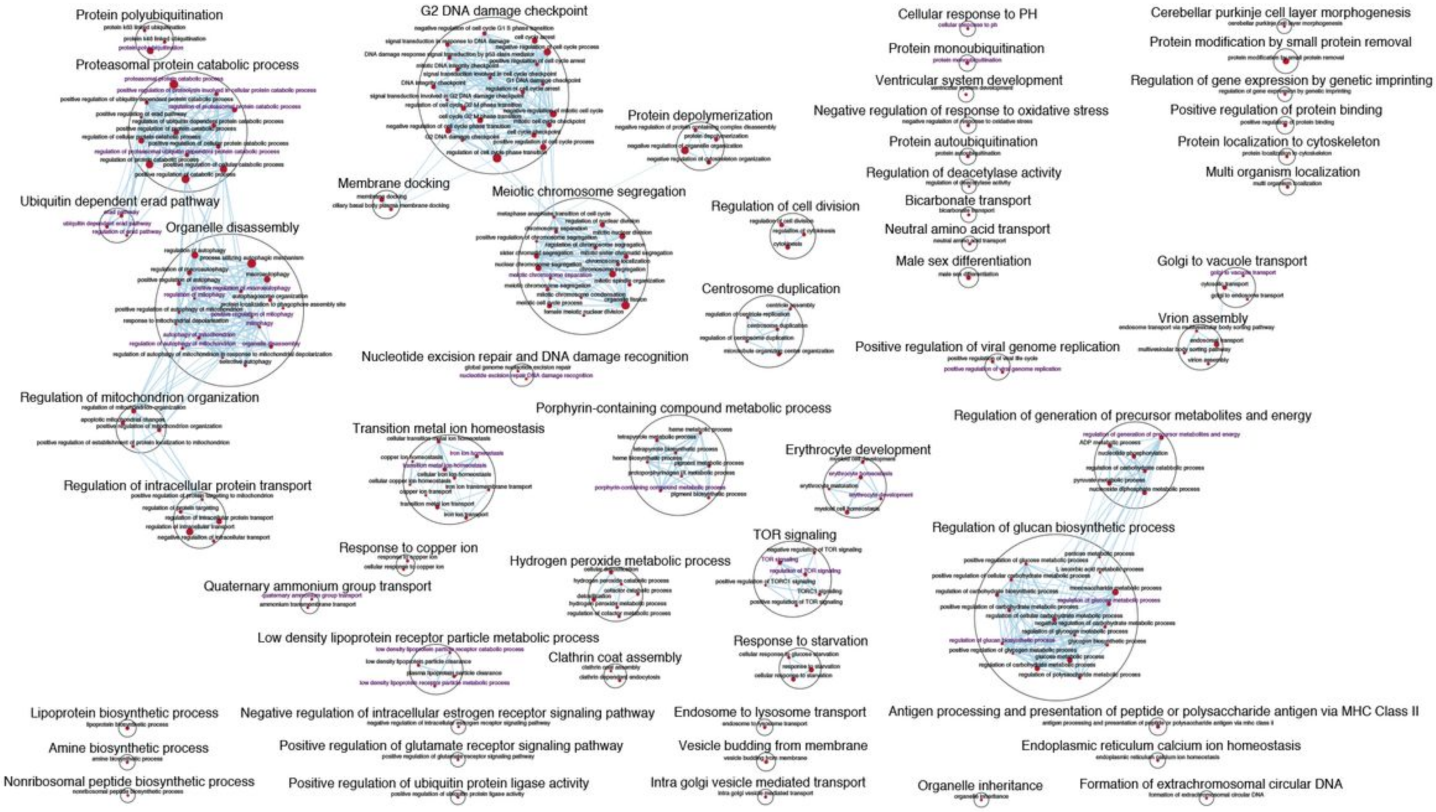

Protein polyubiquitination
Proteasomal protein catabolic process
Ubiquitin dependent erad pathway
Organelle disassembly
Regulation of mitochondrion organization
Regulation of intracellular protein transport
G2 DNA damage checkpoint
Membrane docking
Meiotic chromosome segregation
Protein depolymerization
Regulation of cell division
Centrosome duplication
Nucleotide excision repair and DNA damage recognition
Cellular response to PH
Protein monoubiquitination
Ventricular system development
Negative regulation of response to oxidative stress
Protein autoubiquitination
Regulation of deacetylase activity
Bicarbonate transport
Neutral amino acid transport
Male sex differentiation
Positive regulation of viral genome replication
Cerebellar purkinje cell layer morphogenesis
Protein modification by small protein removal
Regulation of gene expression by genetic imprinting
Positive regulation of protein binding
Protein localization to cytoskeleton
Multi organism localization
Golgi to vacuole transport
Vrion assembly
Porphyrin-containing compound metabolic process
Transition metal ion homeostasis
Erythrocyte development
Regulation of generation of precursor metabolites and energy
Response to copper ion
Quaternary ammonium group transport
Hydrogen peroxide metabolic process
TOR signaling
Regulation of glucan biosynthetic process
Low density lipoprotein receptor particle metabolic process
Clathrin coat assembly
Response to starvation
Lipoprotein biosynthetic process
Amine biosynthetic process
Nonribosomal peptide biosynthetic process
Negative regulation of intracellular estrogen receptor signaling pathway
Positive regulation of glutamate receptor signaling pathway
Positive regulation of ubiquitin protein ligase activity
Endosome to lysosome transport
Vesicle budding from membrane
Intra golgi vesicle mediated transport
Antigen processing and presentation of peptide or polysaccharide antigen via MHC Class II
Endoplasmic reticulum calcium ion homeostasis
Organelle inheritance
Formation of extrachromosomal circular DNA

Figure 3

Biological network of the MGI RNA-seq dataset following Gene Ontology (biological process) gene set enrichment analysis in GSEA (v4.0.3) and visualisation in Cytoscape (3.8.0) 40. Each circle (node) represents a gene set and two nodes are connected by lines (edges) indicating shared genes. The size of a node and width of an edge are proportional to the number of genes enriched in a gene set and the number of genes shared between gene sets, respectively. Gene sets that are similar were annotated and clustered to form a biological theme using the AutoAnnotate App 39 in Cytoscape. The most significantly enriched gene set is used to label a gene set cluster, defined by NES. Red node: gene set enriched in EPO4. Purple node label: top gene sets with NES > 1.90. The enrichment map was created with pathway FDR < 0.1, nominal P < 0.05 and Jaccard Overlap coefficient > 0.375 with combined constant k = 0.5.

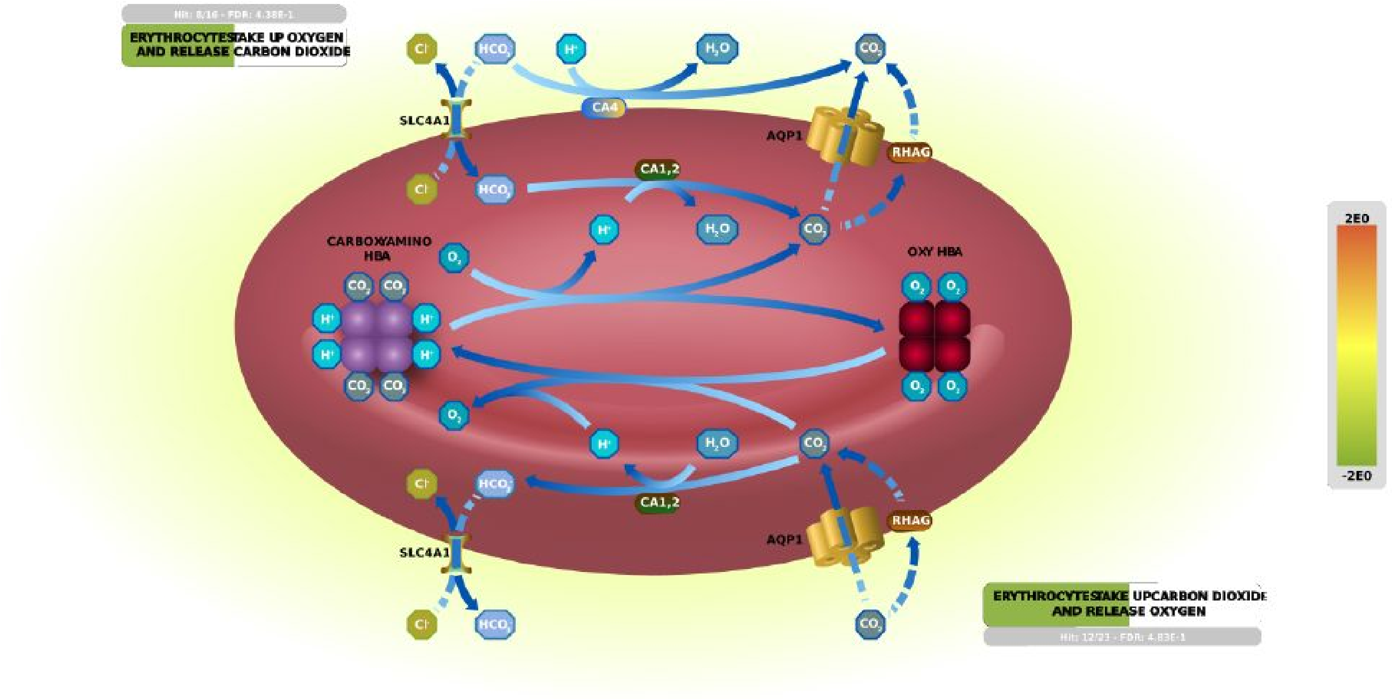

Hit: 8/16 - FDR: 4.38E-1
ERYTHROCYTESTAKE UP OXYGEN AND RELEASE CARBON DIOXIDE
Cl⁻
HCO₃⁻
H⁺
H₂O
CO₂
CA4
SLC4A1
AQP1
RHAG
CA1,2
CARBOXYAMINO HBA
O₂
OXY HBA
2E0
-2E0
ERYTHROCYTESTAKE UPCARBON DIOXIDE AND RELEASE OXYGEN
Hit: 12/23 - FDR: 4.83E-1

Figure 4

Enhanced high-level Reactome pathway diagram for O2/CO2 exchange in erythrocytes 48 by expression overlay with the MGI RNA-seq Post7 dataset. This high-level diagram represents two subpathways, namely erythrocyte take up oxygen and release carbon dioxide and erythrocyte take up carbon dioxide and release oxygen. The green band indicates the proportion of the pathway that is represented in the MGI RNA-seq Post7 dataset, and the colour (green) represents the down-regulation of the pathway genes. The grey bar contains the information for the number of pathway entities in the query dataset, the total number of the pathway entities, and the FDR corrected over-representation probability.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryData112.zip
- WangetalSupplementaryInformation.pdf