A molecular classification of human mesenchymal stromal cells

Florian Rohart, Elizabeth Mason, Nicholas Matigian, Rowland Mosbergen, Othmar Korn, Tyrone Chen, Suzanne Butcher, Jatin Patel, Kerry Atkinson, Kiarash Khosrotehrani, Nicholas Fisk, Kim-Anh Lê Cao, Christine A Wells

Mesenchymal Stromal Cells (MSC) are widely used for the study of mesenchymal tissue repair, and increasingly adopted for cell therapy, despite the lack of consensus on the identity of these cells. In part this is due to the lack of specificity of MSC markers. Distinguishing MSC from other stromal cells such as fibroblasts is particularly difficult using standard analysis of surface proteins, and there is an urgent need for improved classification approaches. Transcriptome profiling is commonly used to describe and compare different cell types, however efforts to identify specific markers of rare cellular subsets may be confounded by the small sample sizes of most studies. Consequently, it is difficult to derive reproducible, and therefore useful markers. We addressed the question of MSC classification with a large integrative analysis of many public MSC datasets. We derived a sparse classifier (The Rohart MSC test) that accurately distinguished MSC from nonMSC samples with >97% accuracy on an internal training set of 635 samples from 41 studies derived on 10 different microarray platforms. The classifier was validated on an external test set of 1291 samples from 65 studies derived on 15 different platforms, with >95% accuracy. The genes that contribute to the MSC classifier formed a protein-interaction network that included known MSC markers. Further evidence of the relevance of this new MSC panel came from the high number of Mendelian disorders associated with mutations in more than 65% of the network. These result in mesenchymal defects, particularly impacting on skeletal growth and function. The Rohart MSC test is a simple *in silico* test that accurately discriminates MSC from fibroblasts, other adult stem/progenitor cell types or differentiated stromal cells. It has been implemented in the www.stemformatics.org resource, to assist researchers wishing to benchmark their own MSC datasets or data from the public domain. The code is available from the CRAN repository and all data used to generate the MSC test is available to download via the Gene Expression Omnibus or the Stemformatics resource.

A Molecular Classification of Human Mesenchymal Stromal Cells

Florian Rohart[1], Elizabeth Mason[1], Nicholas Matigian[1], Rowland Mosbergen[1], Othmar Korn[1], Tyrone Chen[1], Suzanne Butcher[1], Jatin Patel[2], Kerry Atkinson[2], Kiarash Khosrotehrani[2,3], Nicholas M Fisk[2,4], Kim-Anh Lê Cao[3] and Christine A Wells[1*]

[1] Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, Brisbane, QLD Australia 4072

[2] The University of Queensland Centre for Clinical Research, Herston, Brisbane, Queensland, Australia, 4029

[3] The University of Queensland Diamantina Institute, Translational Research Institute, Woolloongabba, Brisbane QLD Australia, 4102

[4] Centre for Advanced Prenatal Care, Royal Brisbane & Women's Hospital, Herston, Brisbane, Queensland, Australia, 4029

*Correspondence to: Christine Wells, c.wells@uq.edu.au

Introduction

Adult tissues maintain the capacity to be replenished as part of the normal processes of homeostasis and repair. The adult stem cell hypothesis proposes that multipotent cells resident in tissues are the source of this cellular renewal, and expand in response to tissue injury. MSC were first isolated from bone marrow, where these occupy an important stem cell niche required for reconstitution of bone and the stromal compartments of marrow, and also play a supportive role in haematopoiesis (*1*, *2*). Subsequently adult stromal progenitors have been isolated and cultured from most organs including placenta, heart, adipose tissue and kidneys although the identity of these cells remains controversial (reviewed by (*3*, *4*)). Specifically the question of how similar cells isolated outside the bone marrow niche are, and whether these could be considered bona fide MSC, or indeed challengingly, whether MSC isolated from different tissues share any phenotypic or molecular characteristics at all (*3*). In this light various cells described as MSC (whether by name or attribution) have been reported as having quite different self-renewal capacity, immunomodulatory properties or propensity to differentiate *in vivo* (*5*). It has been variously argued that MSC isolated from most stromal tissues are derived from perivascular progenitors (*6*), or recruited from the bone marrow to distal tissue sites (*7*), or that resident stromal progenitors from different tissues must have tissue-restricted phenotypes. The most stringent criteria for MSC are in-vivo, bone forming capacity, however this functional standard is rarely addressed in the majority of MSC studies reported in the literature to date (see for example (*5*, *8*)).

Several groups have attempted to address the demand for improved molecular markers, for example using global proteomics methods(*9*), transcriptome analysis of cells capable of regenerating the bone marrow niche (*10*), or comparison of desirable properties such as capacity to form bone (*8*) and indeed the studies reporting global 'omic analysis of MSC number in the hundreds. Each of these studies identifies a different set of potential markers, but there is little consensus among them. Most human studies have been conducted on very small numbers of donors, so it is difficult to dissect donor-donor heterogeneity from source heterogeneity. Nevertheless, line-line variation is a major contributor to differences in MSC growth and differentiation capacity, and clonal variation is evident even when derived from the same donor bone marrow (*8*, *11*). MSC heterogeneity is further compounded by growth conditions, including the density of culture, the inclusion of serum, or the substrate on which they are grown(*12*). Consequently there is little agreement in the literature on definitive molecular or cellular phenotypes of human cultured MSC, whether from bone marrow or other sources.

There is little consensus on whether MSC from differing tissue sources share common functional attributes. The lack of definitive markers for human MSC is a major barrier to understanding genuine similarities, or resolving differences between various cell sources or subsets. Even if acknowledging that there should be functional differences between MSC isolated from different tissues, or donor groups, it is not clear whether there should be any over-arching commonalities that might indicate shared homeostatic roles or ontogenies. The field requires improved methods for benchmarking MSC cultures, including molecular methods that lack the ambiguity of current antibody-based

methods. Here we describe a sophisticated integrative transcriptome analysis of public MSC datasets, and provide a highly accurate *in silico* tool for straightforward assessment of the identity of an MSC culture.

Material and Methods

Design of test and training datasets

A careful screening of all the datasets collated in www.stemformatics.org (Wells et al., 2013), GEO (Barrett et al., 2011) and ArrayExpress (Parkinson et al., 2011) at the time of this analysis identified 120 possible MSC microarray datasets. These were evaluated for the availability of the primary (unprocessed) data; unambiguous replication (biological not technical); the quality control metrics of RNA quality (5'-3' probe ratios); linear range (box-whisker plots of sample median, min and max absolute and normalized values); unambiguous sample descriptions; and sample clustering concordant with the original publication. 35/120 datasets failed these criteria and were excluded from the study.

As the range of phenotypes employed across the remaining 85 MSC microarray studies was broad, we assigned to the training group only those MSC datasets that met at least the following criteria in common: Adherence, Cell surface markers CD105+, CD73+, CD45- and differentiation to at least 2 of the three MSC-definitive lineages (bone, cartilage or fat). As can be seen in supplementary table S2, all training datasets included substantial phenotyping above these minimal criteria. These minimal common criteria were hard-coded into the Stemformatics annotation pipeline, we had a dedicated annotator responsible for the quality of these annotations and these were reviewed

independently by two additional annotators. Sixteen MSC datasets met our 'gold standard' training set criteria for accompanying phenotype of MSCs, together with 27 datasets containing cells from non-mesenchymal or non-stromal sources, which we refer to as non-MSCs. In total, 41 datasets were included in the training set, with two datasets containing both MSCs and non-MSCs, with a total of 125 MSC samples and 510 non-MSC samples from 10 different microarray platforms (Table S3, accompanies the MSC clustering in Figure 2). The remaining MSC datasets were assigned to the independent test set and were used only for evaluation of accuracy of the final signature.

Details on the samples, datasets and references of the experiments can be found in Tables S2, S3 and S5. Two large datasets – 5003 (211 non-MSCs) and 6063 (45 MSCs), were subsampled prior to assigning to the training set to avoid unbalanced results. The samples left out were included in the test set (Table S5). It consisted of 65 experiments (1291 samples, 213 MSCs and 499 non-MSC) profiled across 15 different platforms.

Pre-processing of the data

All data were processed using the R programming language (R Development Core Team, 2011; Venables and Smith, 2008). The pre-processing step involved a background correction performed with affy(Gautier et al., 2004), oligo(Carvalho and Irizarry, 2010) or lumi(Du et al., 2008) packages for processing of microarray data depending on the platform, a log2 transformation of the raw values and a YuGene transformation(Lê Cao et al., 2014). YuGene is a rescaling method using the cumulative

proportion that is applied per sample rather than per dataset or per series. This is highly advantageous as we performed 10-fold cross-validation that would otherwise require renormalization as datasets were added or removed.

In order to combine all the datasets described in Table S2, probes were mapped to Ensembl gene to provide a common set of identifiers. Mapping thresholds of 98% match were used to align microarray probes to Ensembl human v69 transcript model cDNA and ncRNA sequences obtained from Ensembl. Transcript IDs in resulting mapping were converted to Gene IDs using EnsemblBiomart v69(Zhang et al., 2011). In the case of multi-mapping (several probes mapping to the same Ensembl gene ID), the probe with the highest average expression was chosen, on a per-dataset basis.

The combined training data set included the gene expression measurement of 41,185 genes mapped by at least one probe in one dataset. When a dataset had no probes mapping to a particular gene, the expression values of the gene were arbitrarily set to zero for all samples from that dataset. A pre-screening step was then performed to discard genes that were not present in at least half of the samples.

Identification of the 16-gene signature and assignation of a test sample to the MSC or non-MSC class

The MSC signature was identified using a novel implementation of the sparse variant of Partial Least Square Discriminant Analysis (sPLS-DA) (Barker and Rayens, 2003) implemented for multiple microarray studies using the mixOmics package (Lê Cao et al., 2009, 2011). Full details of the statistical model are provided in the Supplementary methods. The underlying code for the statistical test is available as BootsPLS in the

CRAN repository, and we have also made available the d3 code for the interactive MSC graph implemented in Stemformatics via the BioJS framework at http://biojs.io/d/biojs-vis-rohart-msc-test

Network analysis

Twenty-six genes selected on component 1 equated to 18 proteins with a curated interaction in the Networkanalyst protein interaction database (which draws on the PPI database of the International Molecular Exchange (IMEx) consortium (Orchard et al., 2012; Xia et al., 2014) These seed proteins were annotated to a shortest-path first-order network of 42 nodes and 52 PPI edges. Randomised sets of equivalent size were selected from the background (expressed) genes to demonstrate a lack of PPI structure by chance. Gene ontology analysis was assessed using hypergeometric mean against the Jan 2015 EBI UniProt GO library (Huntley et al., 2015). Disease annotations were undertaken using the OMIM (Baxevanis, 2012) and MGI (Shaw, 2009) databases. Subcellular location annotations were taken from UniProt (EMBL et al., 2013).

Differential expression analysis:

Individual MSC markers were assessed for differential analysis between MSC and non-MSC groups using a standard 2-tailed t-test, with a significance threshold of 10^{-6}. For exploration of MSC subsets, a linear mixed model with dataset as random effect was fitted for each gene for which both the mean of bone marrow samples and other sites were higher than the median of all gene expression values. This retained 16,903 genes.

147 P-values were obtained by ANOVA and corrected for multiple testing with the
148 Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).
149

150 **Results**

151 ***Common MSC markers group MSC from bone marrow and other tissues.***

152 The International Society for Cellular Therapy (Dominici et al., 2006) has collated a
153 large set of markers commonly used to immunophenotype MSC. These were used, in
154 combination with more recently identified markers from the current literature (Lv et al.,
155 2014), to assess whether a transcript-based approach might provide a useful molecular
156 tool to identify MSC populations (Supplementary Table S1). In order to compare data
157 generated on different microarray platforms, we built a PLS-DA matrix using these
158 markers and their corresponding expression in highly verified MSC samples. The
159 resulting scatter plot (PLS-DA, Figure 1A) demonstrated the capacity to distinguish
160 between most MSC and non-MSC samples at a transcriptional level, and further
161 showed that MSC isolated from different tissues do cluster together using these
162 markers. Figure 1B shows the 16 of 32 commonly used MSC markers that were
163 significantly differentially expressed between MSC and non-MSC groups ($P<10^{-6}$), and
164 these included CD73 (NTE5), CD105 (Endoglin), PDGFRB and VCAM1. The average
165 expression of the remaining markers is provided in Supplementary Figure 1. Despite
166 ISCT recommendations, most of the MSC publications reviewed herein used a small
167 subset of these antibodies when phenotyping MSC, and CD73+, CD105+ and CD45-
168 were the most consistent subset used (in combination with additional markers and
169 phenotypic information, Supplementary Table 2). When just these three markers were

used to cluster all of the samples, 85% of MSC still grouped together (12/125 misclassified, Table 1, Figure 1A), but almost 12% of non-MSC samples also clustered with this group. The overall accuracy increased to 92% when all 32 markers were used, but the rate of non-MSC misclassification remained high (7%, 35/510) and the majority of these (73.5%) were fibroblasts. It may be that these markers are less stably detected at a mRNA than protein level, however this high misclassification rate is also consistent with a large body of literature documenting the ambiguity of these markers, which are shared with stromal fibroblasts, endothelial progenitors and hematopoietic cells. The variable expression of all 32 markers (Figure 1B, Supplementary Figure S1) is consistent with the reported variability of marker use in the wider MSC research community (reviewed by (Lv et al., 2014; Samsonraj et al., 2015)). Nevertheless, the capacity of these known markers to cluster MSC from different studies gave us confidence that a transcriptome approach was a useful and simplified alternate to antibody-based protocols, so we next took an unbiased approach to find a set of markers that could improve on the current classification paradigm. Our goal was to find an *in silico* marker set that reproducibly identified *bona fide* MSC samples regardless of platform or laboratory differences, and provide a molecular test that was simpler, and more accurate than current methods.

Derivation of an improved, simple and accurate in silico MSC classifier.

A careful review of the public databases identified 120 potential MSC transcriptome studies, each comprising of a small number of donors. These were carefully curated for source, phenotypic information and growth conditions (see methods for details). From

these efforts, a gold standard 'training set' was identified as meeting high confidence MSC phenotype including at least the minimal common set of CD73+, CD105+, CD45- and bilineage differentiation. The training set consisted of 125 MSC samples from 16 independently derived datasets derived predominantly from bone marrow, but also included studies from other adult, neonatal and fetal stromal sources. MSC were compared to 510 definitively non-MSC samples from primary human tissues and cell lines, including cultured fibroblasts, haematopoietic cells and pluripotent stem cell lines (Supplementary Tables S2, S3).

To fully integrate and interrogate this data, we derived a novel cross-study analysis framework. Our approach, described in Figure 2A, included a cross-platform normalisation step (Lê Cao et al., 2014), and a modified variable (gene) selection methodology. The first part of the protocol identified hundreds of potential MSC markers, which in combination greatly improved the classification accuracy of 97.7% (Table 1). This included many of the known MSC markers. Each gene was further evaluated for stability by subsampling the datasets to ensure that its inclusion was not reliant on one dominant source or platform. Stability is indicated by the probability of selection over 200 iterations in Figure 2B, and was the step that excluded most of the commonly used MSC markers. For example, PDGFRB and VCAM1 were identified as potential component 1 genes but their inclusion was highly variable (0.76 and 0.59 probability of selection respectively).

We reasoned that if the majority of genes discriminating between MSC and non-MSC are describing a common biology and are highly correlated, then a subset of these

genes could be identified that would represent the entire network. Therefore we iteratively assessed how the inclusion of each gene contributed to the overall accuracy of the signature. This found the subset of variables that were most stable and least redundant at a statistical level, and that would represent the greater network of MSC-related measurements (Figure 2C). Sixteen genes were identified, collectively forming a 'signature', which provided a high degree of discrimination between MSC and non-MSC cell types, without any loss of accuracy in accurately identifying MSC (>95% correct MSC call or 4/125 misclassified MSC samples, Table 1) and with improved discrimination from fibroblasts and other non-MSC cell types (1.61% false positive, Table 1). We confirmed that this clustering was agnostic to technology platform or manufacturer (Supplementary Figure S2).

Cells derived from bone marrow were reliably grouped together with this method (Figure 2D, Supplementary Figure S2E), and MSC from other tissue sources, including adipose tissue, skin, lung, placenta and cord blood shared this signature. Each gene in the signature made an additive contribution across 4 vectors (components), such that the absolute expression of any one gene might differ from sample to sample but the combination of gene expression was highly predictive. High expression of component 1 genes was most likely to be a positive predictor of an MSC classification (Figure 2 and Supplementary Figure S3A), as indicated by the correlation of expression of each gene with its component. Note that the components are linear vectors, and so a negative correlation (as for component 1 genes) simply indicates the contribution of the genes to clustering MSC on the positive or negative region of that component. The inclusion of

components 2-4 provided higher discrimination for subsets of MSC and non-MSC, particularly differentiating MSC and fibroblasts derived from various tissues. These latter components included stress-related genes (heat shock proteins) and early indicators of lineage commitment (osteomodulin). Importantly, this multicomponent based approach, in contrast to a typical differential expression analysis, allowed for a common MSC phenotype that is also permissive of tissue-specific differences in the wider MSC gene network.

The implementation in www.stemformatics.org assessed the MSC score across 200 iterative predictions, where a sample must have a 95% pass rate to be classed as an MSC. The distribution of the training sample scores was used to determine high confidence scores (Figure 2E). By using 200 subsamplings of the training set, 200 scores were recorded for each sample, which enabled us to derive an individual 95% Confidence Interval (CI). A sample was assigned to the MSC class if the lower bound of its 95%CI is strictly higher than 0.5169. Similarly, a non-MSC classification is given if the upper bound of the 95% CI was lower than 0.4337. Samples failing to meet these criteria were assigned to an ‘unknown’ category. Accordingly, the four misclassified MSC in the training set included one adult bone marrow MSC sample (predicted 1/200 times as MSC), and the remaining from two fetal studies, the first consisting of 10 week chorionic villi (predicted 29/200 times as MSC) and 12-week chorionic membrane preparation (2/200 MSC predictions), the second from a neonatal lung aspirate (0/200 positive MSC predictions).

The MSC signature genes form a cohesive network implicated in healthy mesenchymal development and function.

To assess possible functional relationships between MSC signature genes, we used a curated set of protein-protein interactions from the BioGrid database using the genes selected from component 1 that showed a high discriminating power between MSC and non-MSC. These formed a network of 43 interacting proteins (Figure 3A). The higher expression of these genes in MSC samples is confirmed in Fig 3B. If the statistical tool had identified a random set of genes, then the network would have little connectivity and there would be no relevant functional annotations. This was confirmed by random subsampling from the background datasets, which failed to form any PPI network. To assess whether the highly connected MSC network also shared any cohesive functional annotations, we examined mutation databases for evidence of human diseases associated with network members. A high proportion of the MSC network (30/43) are represented in Mendelian disorders of mesenchymal development by virtue of their mutation spectrum in facial or musculo-skeletal dysmorphologies in man, or evidence of mesodermal defects in KO mouse models (Supplementary Table S4). These included the paired-related homeobox-1 (*PRRX1*), a transcription factor important for early embryonic skeletal and facial development, and with a *de novo* mutation spectrum in the embryonic dysmorphology syndrome Agnathia-otocephaly (Çelik et al., 2012). Likewise, mutations in bone morphogenetic protein 14 (*BMP14/GDF5*) lead to developmental abnormalities in chondrogenesis and skeletal bone (Degenkolbe et al., 2013). Mutations in *DDR2* cause limb defects, including spondylo-epiphyseal-metaphyseal dysplasia (Ali et al., 2010) and mice over-expressing *DDR2* have

increased body size and atypical body fat (Kawai et al., 2014). In humans, Polymorphisms in *ABI3BP* are associated with increased risk of osteochondropathy (Zhang et al., 2014), and mice lacking Abi3bp have profound defects in MSC differentiation to bone and fat (Hodgkinson et al., 2013).

We next examined functions that had been specifically validated in MSC biology, specifically, whether any members of the signature had been used to prospectively isolate MSC from tissue sources. *ITGA11* was a member of the core signature that has been used to prospectively enrich MSC from bone marrow with enhanced colony forming capacity (Kaltz et al., 2010), and independently shown to be enriched more than 3 fold at protein level in bone marrow MSC compared to dermal fibroblasts or perivascular cells (Holley et al., 2015). Although several of the known and commonly used MSC markers were indeed captured in the large initial set of potential classifiers, but rejected by our statistical method on the grounds of poor selection stability, these were 'rescued' in the protein interaction network. That is, the behavior of these markers was variable across laboratories and between microarray platforms, and often high expressed on non-MSC cell types. Nevertheless, the interaction network demonstrated some cohesive biology with these known markers. The most highly connected member of the extended network was VCAM1, which was identified in the large prospective marker set but with a low frequency of selection (0.6 on component 1), which eliminated it from the final classifier. VCAM1, together with STRO-1, has been used for the prospective isolation of human bone marrow MSC (Gronthos, 2003). VCAM1 is an adhesion molecule that is induced by inflammatory stimuli to regulate leukocyte

adhesion to the endothelium (Dansky et al., 2001); however, in cardiac precursors its expression demarcates commitment to mesenchymal rather than endothelial lineages (Skelton et al., 2014).

Other members of our network that have been previously described in human or mouse MSC biology, and used to prospectively isolate cells or have been validated at the protein level include *PDGFRβ* (Koide et al., 2007), *SPINT2* (Roversi et al., 2014), *CCDC80* (Charbord et al., 2015), *FAP* (Bae et al., 2008), *BGN* (Holley et al., 2015), and *TM4SF1* (Bae et al., 2011). SPINT2 is a serine protease inhibitor whose activity is required in bone-marrow MSC, and its loss alters hematopoietic stem cell function in myelo-dysplastic disorders (Roversi et al., 2014). In mouse, CCDC80 is also necessary for reconstitution of bone marrow and support of haematopoiesis (Charbord et al., 2015).

The network included a high proportion of extracellular proteins (54%) with demonstrated roles in the modification of extracellular matrix proteins including proteoglycans, as well as regulators of growth factor and cytokine signalling. This included the cell migration inducing protein (KIAA1199/ CEMIP), which is secreted in its mature form. It regulates Wnt and TGFβ3 signalling by depolarising hyaluronan, and may alter trafficking of cytokines and growth factors to the extracellular milieu (Yoshida et al., 2013). DDR2 is a receptor tyrosine kinase that interacts directly with collagens. It stabilises the transcription factor SNAIL, and has been implicated in epithelial-mesenchyme transitions in epithelial cancers (Zhang et al., 2013). CCDC80 binds syndecan-heparin sulphate containing proteoglycans, has been shown to inhibit

WNT/beta-catenin signalling and has a regulatory role in adipogenesis (Tremblay et al., 2009; Walczak et al., 2014). SRPX2 is a secreted chondroitin sulfate proteoglycan involved in endothelial cell migration, tissue remodelling and vascular sprouting (Royer-Zemmour et al., 2008). The chaperonins HSPB5/CRYAB and HSPB6 stabilise protein complexes, and may assist in delivery of growth factor complexes where these are present in high concentrations. In transplantation paradigms it is likely that the therapeutic benefit derived from MSC is via local immunomodulatory, anti-inflammatory, and/or trophic effects during the acute phase of cell therapy. The network of genes identified here as enriched in MSC suggests an over-arching role for these cells in modifying the extracellular environment, functions important in development as well as in homeostatic regulation of adult tissues.

MSC differentiation, dedifferentiation and the MSC signature

The majority of public microarray datasets available to us had limited phenotypic data available, so these were not used to derive our MSC signature. Nevertheless we annotated each of these samples as *presumptive* MSC (213 samples) or *presumptive* non-MSC (499 samples) based on their origin and use in the source publication (Supplementary Table S5). Where MSC were profiled during *in vitro* lineage differentiation, we assigned the samples taken at intermediate time points to an 'unknown' category (579 samples) prior to testing these with the signature. Implementation of the Rohart Test in the www.stemformatics.org resource allowed us to evaluate a wide range of different experimental paradigms. Despite the lack of phenotypic information associated with these datasets, the agreement between

publication status and our classification was high. Five percent of the presumptive non-MSC (27/499) were misclassified by the signature as MSC, and around half of these (>13) were neonatal or fetal dermal fibroblasts (Supplementary Table S5. Others have reported MSC fractions derived from dermal tissues (reviewed in (Vaculik et al., 2012)) and certainly fibroblasts from other sources were not classified as MSC. Furthermore, the signature could discriminate between MSC and differentiating cultures. Figure 3C demonstrates loss of the MSC score during chondrogenic differentiation with the addition of TGFβ (Dataset 6119 (Mrugala et al., 2009)) and this pattern was recapitulated for cells differentiating to mineralising bone or to adipose-like cells or when undergoing reprogramming of an adipose-tissue derived iPSC (data not shown, but available in the Stemformatics resource).

Comparison of MSC and adult stem/progenitor cell types

The limbal cell niche hosts both limbal epithelial and stromal progenitors (Lim et al., 2012), and the stromal progenitors were also classified as MSC by our tool (Dataset 6450). Some MSC subsets are likely to be derived from a perivascular progenitor. In our hands, primary skeletal-muscle mesoangioblasts thought to be a subset of perivascular cells in skeletal and smooth muscle (Dataset 6265 (Tedesco et al., 2012), defined as alkaline-phosphatase^{+} CD146^{+} CD31/Epcam^{-} CD56/Ncam^{-} with demonstrated skeletal muscle differentiation, were classified as MSC (Figure 3D). In contrast, the majority of cells derived from a perivascular location (and confirmed as such with tissue imaging in the source publication) were not classified as MSC (Figure 3E). On examining putative markers of perivascular progenitors in these samples, we could demonstrate that the

majority of perivascular progenitors expressed higher levels of Nestin than the majority of MSC (Figure 3F). MCAM+ and MCAM- cells were apparent in both MSC and pericyte groups, although a higher proportion of perivascular progenitor expressed MCAM RNA. In contrast, PDGFRA was highly expressed in MSC but not informative in perivascular cells, and PDGFRB was highly expressed in both populations. Others have shown that high expression of PDGFRA is associated with highly proliferative MSC colonies, suggesting that its expression is associated with expansion in culture (Samsonraj et al., 2015). These data are consistent with a classification hierarchy determined by mouse and human lineage studies, where multipotent adult cells are quiescent in a perivascular location (Acar et al., 2015; Crisan et al., 2008). Thus perivascular progenitor cells with MSC differentiation capacity are defined as Rohart test negative, Nestin positive in our test, and as such are distinct from a Rohart test positive MSC. Cells differentiating to osteoblast, chondrocyte, adipocyte or fibroblast exit the MSC state and rapidly become negative for the Rohart MSC score. Given that a proportion of Rohart test positive MSC express MCAM or Nestin, the classification tool may detect a phenotypic spectrum that spans the intermediates across the perivascular-MSC-fibroblast hierarchy.

Tssue clustering of MSC is confounded by gender and MHC-1 haplotype.

The capacity to group MSC-like cells is consistent with the general assumption that MSC from different tissue share some common molecular properties. Many of the individual studies in this reanalysis describe tissue-specific differences in MSC populations. We were not able to recapitulate any of these specific differences on the

integrated dataset. Nevertheless, MSC from different tissues did form subclusters (Supplementary Figures S2, S3), and the majority of bone marrow MSC clustered together (Figure S2E). We therefore examined more broadly the genes that were significantly different between bone marrow MSC and other cell types at the whole transcriptome level. This analysis confirmed the observed clustering of bone marrow derived MSC, distinguished by differential expression of 425 genes (adjusted $P<0.01$, Supplementary Table S6). The genes that were most differentially expressed between the different MSC sources in our combined analysis were MHC class I genes, and these accounted for >40% of the top 100 differentially expressed genes in the bone-marrow comparisons (Supplementary Table S6). The HLA isotypes were generally, but not exclusively, expressed at lower levels in bone marrow MSC (Hierarchical Cluster, Supplementary Figure S3). Estrogen and progesterone receptors, and a network of associated target genes were also significantly different between tissue sources (Supplementary Table S6), and this may reflect a gender bias in tissue sampling; although the gender of the donors was not available for a majority of samples, some tissues (such as decidual sources) will be entirely female in origin. Further molecular sub-classifications of MSC will therefore require much larger studies that address specific clinical or differentiation properties of the cells, and must also consider ascertainment biases that may introduce confounding variables such as HLA subtypes or gender.

Discussion

Modern molecular classification tools are needed for the characterisation of MSC *ex vivo* and *in vivo*. Antibody based methods currently rely on a subset of cell surface proteins that are widely acknowledged to lack specificity, and the reliability of these assays is dependant on operator expertise. Our study set out to provide an alternate test that had better discrimination power than current assays, was robust and easy to generate. In doing so we developed a specific gene signature that is shared by a wide-variety of MSC. The “Rohart MSC test” is an *in silico* tool that has been implemented as a simple online test that will be useful in standardisation or improvement of current bulk isolation methods. This classification tool is available in the Stemformatics.org platform, together with all the primary data used in derivation of the signature. Details on submitting proprietary data to the Rohart test are available on the stemformatics.org site.

All together we curated more than 120 MSC-related gene expression datasets in the www.stemformatics.org resource (Wells et al., 2012); the datasets can be queried here using key word, dataset ID or author, together with an implementation of the Rohart MSC test.

Our approach highlights the potential robustness of biological signatures when combining data from many different sources, where experimental variables such as platform or batch can be reduced (Figure S2). The methods we used for derivation of a common MSC classifier could be applied to the meta-analysis of any cell subset or phenotype where sufficient samples can be drawn from public expression databases.

The Rohart test provides a snap shot of the current state of play in MSC biology. As an *in silico* test it reflects all of the ambiguities existing in current nomenclature and culture practise. We anticipate that a computational classifier will evolve as the field of MSC biology evolves, and as isolation methods improve. Indeed, the question of what is an MSC, and whether these are a distinct stem cell population recruited from the bone marrow, as suggested by mouse studies of fetomaternal microchimerism (Seppanen et al., 2013) or from perivasculature, as suggested by immunotagging of MSC-like cells from perivascular regions in human tissues (Crisan et al., 2008), or are resident progenitor populations specific to each organ cannot be resolved in the current study. The signature itself is dependent on the quality of the MSC used in the training set. As rare adult stem/progenitor cell types were under-represented in the current test or training datasets, we anticipate that functional classification of MSC subtypes will improve as newer sampling methods provide the means to identify and replicate these cells. To highlight this point, the signature distinguishes perivascular progenitors from MSC, however resolving a perivascular progenitor signature would require substantially more data on this population than is currently available in the public domain. We expect that further refinements in the isolation or culture of purer MSC or more precisely defined functional subsets will also result in future evolutions of this *in silico* signature.

In summary, we set out to systematically review the current state of play in MSC biology using a meta-analysis of transcriptome studies, and in doing so were able robustly to identify a general MSC phenotype that could distinguish MSC from other cell types. The

resulting signature could also identify points of transition as MSC underwent differentiation or reprogramming studies. Furthermore, we demonstrated that, at least at a gene expression level, our *de novo* derived signature outperformed the classification accuracy of the combined set of traditional MSC cell surface markers. While a signature approach such as ours is not able to resolve the ontogeny or in vivo function of MSC, it does provide a tool for better benchmarking and comparison of the cells grown ex vivo, and will assist with comparison of cells derived for clinical purposes. The methods that we describe here, and the resulting molecular classifier represent an important step towards addressing the more intractable questions of MSC identity, ontogenic relationships and function.

References

Acar, M., Kocherlakota, K.S., Murphy, M.M., Peyer, J.G., Oguro, H., Inra, C.N., Jaiyeola, C., Zhao, Z., Luby-Phelps, K., and Morrison, S.J. (2015). Deep imaging of bone marrow shows non-dividing stem cells are mainly perisinusoidal. Nature *526*, 126–130.

Ali, B.R., Xu, H., Akawi, N.A., John, A., Karuvantevida, N.S., Langer, R., Al-Gazali, L., and Leitinger, B. (2010). Trafficking defects and loss of ligand binding are the underlying causes of all reported DDR2 missense mutations found in SMED-SL patients. Hum. Mol. Genet. *19*, 2239–2250.

Bae, S., Park, C.W., Son, H.K., Ju, H.K., Paik, D., Jeon, C.-J., Koh, G.Y., Kim, J., and Kim, H. (2008). Fibroblast activation protein alpha identifies mesenchymal stromal cells from human bone marrow. Br. J. Haematol. *142*, 827–830.

Bae, S., Shim, S.H., Park, C.W., Son, H.K., Lee, H.J., Son, J.Y., Jeon, C., and Kim, H. (2011). Combined omics analysis identifies transmembrane 4 L6 family member 1 as a surface protein marker specific to human mesenchymal stem cells. Stem Cells Dev. *20*, 197–203.

Barker, M., and Rayens, W. (2003). Partial least squares for discrimination. J. Chemom. *17*, 166–173.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., et al. (2011). NCBI GEO: archive for functional genomics data sets—10 years on. Nucleic Acids Res *39*, D1005–D1010.

Baxevanis, A.D. (2012). Searching online mendelian inheritance in man (OMIM) for information on genetic loci involved in human disease. Curr. Protoc. Bioinforma.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B *57*, 289–300.

Bianco, P., Cao, X., Frenette, P.S., Mao, J.J., Robey, P.G., Simmons, P.J., and Wang, C.Y. (2013). The meaning, the sense and the significance: translating the science of mesenchymal stem cells into medicine. Nat Med *19*, 35–42.

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

Çelik, T., Simsek, P.O., Sozen, T., Ozyuncu, O., Utine, G.E., Talim, B., Yiğit, Ş., Boduroglu, K., and Kamnasaran, D. (2012). PRRX1 is mutated in an otocephalic newborn infant conceived by consanguineous parents. Clin. Genet. *81*, 294–297.

Charbord, P., Pouget, C., Binder, H., Dumont, F., Stik, G., Levy, P., Allain, F., Marchal, C., Richter, J., Uzan, B., et al. (2015). A Systems Biology Approach for Defining the Molecular Framework of the Hematopoietic Stem Cell Niche. Cell Stem Cell *15*, 376–391.

Crisan, M., Yap, S., Casteilla, L., Chen, C.-W., Corselli, M., Park, T.S., Andriolo, G., Sun, B., Zheng, B., Zhang, L., et al. (2008). A Perivascular Origin for Mesenchymal Stem Cells in

517 Multiple Human Organs. Cell Stem Cell *3*, 301–313.

518 Dansky, H.M., Barlow, C.B., Lominska, C., Sikes, J.L., Kao, C., Weinsaft, J., Cybulsky, M.I., and
519 Smith, J.D. (2001). Adhesion of monocytes to arterial endothelium and initiation of
520 atherosclerosis are critically dependent on vascular cell adhesion molecule-1 gene dosage.
521 Arterioscler. Thromb. Vasc. Biol. *21*, 1662–1667.

522 Degenkolbe, E., König, J., Zimmer, J., Walther, M., Reißner, C., Nickel, J., Plöger, F.,
523 Raspopovic, J., Sharpe, J., Dathe, K., et al. (2013). A GDF5 Point Mutation Strikes Twice -
524 Causing BDA1 and SYNS2. PLoS Genet *9*, e1003846.

525 Dominici, M., Le Blanc, K., Mueller, I., Slaper-Cortenbach, I., Marini, F., Krause, D., Deans, R.,
526 Keating, A., Prockop, D., and Horwitz, E. (2006). Minimal criteria for defining multipotent
527 mesenchymal stromal cells. The International Society for Cellular Therapy position statement.
528 Cytotherapy *8*, 315–317.

529 Du, P., Kibbe, W.A., and Lin, S.M. (2008). lumi: a pipeline for processing Illumina microarray.
530 Bioinformatics *24*, 1547–1548.

531 EMBL, SIB Swiss Institute of Bioinformatics, and Protein Information Resource (PIR) (2013).
532 UniProt. In Nucleic Acids Research, p. 41: D43–D47.

533 Friedenstein, A.J., Piatetzky-Shapiro, I.I., and Petrakova, K. V (1966). Osteogenesis in
534 transplants of bone marrow cells. J. Embryol. Exp. Morphol. *16* , 381–390.

535 Gautier, L., Cope, L., Bolstad, B.M., and Irizarry, R.A. (2004). affy--analysis of Affymetrix
536 GeneChip data at the probe level. Bioinformatics *20*, 307–315.

537 Gronthos, S. (2003). Molecular and cellular characterisation of highly purified stromal stem cells
538 derived from human bone marrow. J. Cell Sci. *116*, 1827–1835.

539 Hodgkinson, C.P., Naidoo, V., Patti, K.G., Gomez, J.A., Schmeckpeper, J., Zhang, Z., Davis, B.,
540 Pratt, R.E., Mirotsou, M., and Dzau, V.J. (2013). Abi3bp is a multifunctional autocrine/paracrine
541 factor that regulates mesenchymal stem cell biology. Stem Cells *31*, 1669–1682.

542 Holley, R.J., Tai, G., Williamson, A.J.K., Taylor, S., Cain, S.A., Richardson, S.M., Merry, C.L.R.,
543 Whetton, A.D., Kielty, C.M., and Canfield, A.E. (2015). Comparative Quantification of the
544 Surfaceome of Human Multipotent Mesenchymal Progenitor Cells. Stem Cell Reports.

545 Huntley, R.P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M.J., and
546 O'Donovan, C. (2015). The GOA database: Gene Ontology annotation updates for 2015.
547 Nucleic Acids Res. *43* , D1057–D1063.

548 Kaltz, N., Ringe, J., Holzwarth, C., Charbord, P., Niemeyer, M., Jacobs, V.R., Peschel, C.,
549 Häupl, T., and Oostendorp, R.A.J. (2010). Novel markers of mesenchymal stem cells defined by
550 genome-wide gene expression analysis of stromal cells from different sources. Exp. Cell Res.
551 *316*, 2609–2617.

552 Kawai, I., Matsumura, H., Fujii, W., Naito, K., Kusakabe, K., Kiso, Y., and Kano, K. (2014).

553 Discoidin domain receptor 2 (DDR2) regulates body size and fat metabolism in mice.
554 Transgenic Res. *23*, 165–175.

555 Koide, Y., Morikawa, S., Mabuchi, Y., Muguruma, Y., Hiratsu, E., Hasegawa, K., Kobayashi, M.,
556 Ando, K., Kinjo, K., Okano, H., et al. (2007). Two distinct stem cell lineages in murine bone
557 marrow. Stem Cells *25*, 1213–1221.

558 Lê Cao, K.-A., Rohart, F., McHugh, L., Korn, O., and Wells, C.A. (2014). YuGene: A simple
559 approach to scale gene expression data derived from different platforms for integrated analyses.
560 Genomics *103*, 239–251.

561 Lê Cao, K.A., Martin, P., Robert-Granié, C., and Besse, P. (2009). Sparse canonical methods
562 for biological data integration: application to a cross-platform study. BMC Bioinformatics *10*, 34.

563 Lê Cao, K.A., Boitard, S., and Philippe, B. (2011). Sparse PLS discriminant analysis: biologically
564 relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics
565 *12*.

566 Lee, E.S.M., Bou-Gharios, G., Seppanen, E., Khosrotehrani, K., and Fisk, N.M. (2010). Fetal
567 stem cell microchimerism: natural-born healers or killers? Mol. Hum. Reprod. *16* , 869–878.

568 Li, G., Zhang, X.A., Wang, H., Wang, X., Meng, C.L., Chan, C.Y., Yew, D.T.W., Tsang, K.S., Li,
569 K., Tsai, S.N., et al. (2009). Comparative proteomic analysis of mesenchymal stem cells derived
570 from human bone marrow, umbilical cord, and placenta: Implication in the migration. Proteomics
571 *9*, 20–30.

572 Lim, M.N., Hussin, N.H., Othman, A., Umapathy, T., Baharuddin, P., Jamal, R., and Zakaria, Z.
573 (2012). Ex vivo expanded SSEA-4+ human limbal stromal cells are multipotent and do not
574 express other embryonic stem cell markers. Mol. Vis. *18*, 1289–1300.

575 Liu, Y., Muñoz, N., Bunnell, B.A., Logan, T.M., and Ma, T. (2015). Density-Dependent Metabolic
576 Heterogeneity in Human Mesenchymal Stem Cells. Stem Cells n/a – n/a.

577 Lv, F.-J., Tuan, R.S., Cheung, K.M.C., and Leung, V.Y.L. (2014). Concise review: the surface
578 markers and identity of human mesenchymal stem cells. Stem Cells *32*, 1408–1419.

579 Mrugala, D., Dossat, N., Ringe, J., Delorme, B., Coffy, A., Bony, C., Charbord, P., Haupl, T.,
580 Daures, J.P., Noel, D., et al. (2009). Gene expression profile of multipotent mesenchymal
581 stromal cells: Identification of pathways common to TGFbeta3/BMP2-induced chondrogenesis.
582 Cloning Stem Cells *11*, 61–76.

583 Orchard, S., Kerrien, S., Abbani, S., Aranda, B., Bhate, J., Bidwell, S., Bridge, A., Briganti, L.,
584 Brinkman, F.S.L., Cesareni, G., et al. (2012). Protein interaction data curation: the International
585 Molecular Exchange (IMEx) consortium. Nat Meth *9*, 345–350.

586 Parkinson, H., Sarkans, U., Kolesnikov, N., Abeygunawardena, N., Burdett, T., Dylag, M.,
587 Emam, I., Farne, A., Hastings, E., Holloway, E., et al. (2011). ArrayExpress update—an archive
588 of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic
589 Acids Res *39*, D1002–D1004.

590 Phinney, D.G. (2012). Functional heterogeneity of mesenchymal stem cells: Implications for cell
591 therapy. J Cell Biochem *113*, 2806–2812.

592 Pittenger, M.F., Mackay, A.M., Beck, S.C., Jaiswal, R.K., Douglas, R., Mosca, J.D., Moorman,
593 M.A., Simonetti, D.W., Craig, S., and Marshak, D.R. (1999). Multilineage potential of adult
594 human mesenchymal stem cells. Science (80-.). *284*, 143–147.

595 R Development Core Team, R. (2011). R: A Language and Environment for Statistical
596 Computing. R Found. Stat. Comput. *1*, 409.

597 Reinisch, A., Etchart, N., Thomas, D., Hofmann, N.A., Fruehwirth, M., Sinha, S., Chan, C.K.,
598 Senarath-Yapa, K., Seo, E.-Y., Wearda, T., et al. (2014). Epigenetic and in vivo comparison of
599 diverse MSC sources reveals an endochondral signature for human hematopoietic niche
600 formation. Blood *125*, 249–260.

601 Roversi, F.M., Lopes, M.R., Machado-Neto, J.A., Longhini, A.L.F., Duarte, A. da S.S., Baratti,
602 M.O., Palodetto, B., Corrocher, F.A., Pericole, F.V., Campos, P. de M., et al. (2014). Serine
603 protease inhibitor kunitz-type 2 is downregulated in myelodysplastic syndromes and modulates
604 cell-cell adhesion. Stem Cells Dev. *23*, 1109–1120.

605 Royer-Zemmour, B., Ponsole-Lenfant, M., Gara, H., Roll, P., Lévêque, C., Massacrier, A.,
606 Ferracci, G., Cillario, J., Robaglia-Schlupp, A., Vincentelli, R., et al. (2008). Epileptic and
607 developmental disorders of the speech cortex: Ligand/ receptor interaction of wild-type and
608 mutant SRPX2 with the plasminogen activator receptor uPAR. Hum. Mol. Genet. *17*, 3617–
609 3630.

610 Samsonraj, R.M., Rai, B., Sathiyanathan, P., Puan, K.J., Rötzschke, O., Hui, J.H., Raghunath,
611 M., Stanton, L.W., Nurcombe, V., and Cool, S.M. (2015). Establishing Criteria for Human
612 Mesenchymal Stem Cell Potency. Stem Cells *33*, 1878–1891.

613 Seppanen, E., Roy, E., Ellis, R., Bou-Gharios, G., Fisk, N.M., and Khosrotehrani, K. (2013).
614 Distant mesenchymal progenitors contribute to skin wound healing and produce collagen:
615 evidence from a murine fetal microchimerism model. PLoS One *8*, e62662.

616 Shaw, D.R. (2009). Searching the Mouse Genome Informatics (MGI) resources for information
617 on mouse biology from genotype to phenotype. Curr. Protoc. Bioinformatics *Chapter 1*, Unit1.7.

618 Skelton, R.J.P., Costa, M., Anderson, D.J., Bruveris, F., Finnin, B.W., Koutsis, K., Arasaratnam,
619 D., White, A.J., Rafii, A., Ng, E.S., et al. (2014). SIRPA, VCAM1 and CD34 identify discrete
620 lineages during early human cardiovascular development. Stem Cell Res. *13*, 172–179.

621 Sworder, B.J., Yoshizawa, S., Mishra, P.J., Cherman, N., Kuznetsov, S.A., Merlino, G.,
622 Balakumaran, A., and Robey, P.G. (2015). Molecular profile of clonal strains of human skeletal
623 stem/progenitor cells with different potencies. Stem Cell Res. *14*, 297–306.

624 Tedesco, F.S., Gerli, M.F.M., Perani, L., Benedetti, S., Ungaro, F., Cassano, M., Antonini, S.,
625 Tagliafico, E., Artusi, V., Longa, E., et al. (2012). Transplantation of genetically corrected human
626 iPSC-derived progenitors in mice with limb-girdle muscular dystrophy. Sci. Transl. Med. *4*,

140ra89.

Tremblay, F., Revett, T., Huard, C., Zhang, Y., Tobin, J.F., Martinez, R. V., and Gimeno, R.E. (2009). Bidirectional modulation of adipogenesis by the secreted protein Ccdc80/DRO1/URB. J. Biol. Chem. *284*, 8136–8147.

Vaculik, C., Schuster, C., Bauer, W., Iram, N., Pfisterer, K., Kramer, G., Reinisch, A., Strunk, D., and Elbe-Burger, A. (2012). Human Dermis Harbors Distinct Mesenchymal Stromal Cell Subsets. J Invest Dermatol *132*, 563–574.

Venables, W.N., and Smith, D.M. (2008). R Development Core Team. An Introd. to R Notes R A Program. Environ. Data Anal. Graph. R Core Team Version *2*, R: A language and environment for statistical comp.

Walczak, E.M., Kuick, R., Finco, I., Bohin, N., Hrycaj, S.M., Wellik, D.M., and Hammer, G.D. (2014). Wnt-Signaling Inhibits Adrenal Steroidogenesis by Cell-Autonomous and Non-Cell-Autonomous Mechanisms. Mol. Endocrinol. me20141060.

Wells, C., Mosbergen, R., Korn, O., Choi, J., Seidenman, N., Matigian, N.A., Vitale, A.M., and Shepherd, J. (2012). Stemformatics: Visualisation and sharing of stem cell gene expression. Stem Cell Res. *10*, 387–395.

Wells, C.A., Mosbergen, R., Korn, O., Choi, J., Seidenman, N., Matigian, N.A., Vitale, A.M., and Shepherd, J. (2013). Stemformatics: Visualisation and sharing of stem cell gene expression. Stem Cell Res. *10*, 387–395.

Xia, J., Benner, M.J., and Hancock, R.E.W. (2014). NetworkAnalyst - integrative approaches for protein–protein interaction network analysis and visual exploration. Nucleic Acids Res. *42* , W167–W174.

Yoshida, H., Nagaoka, A., Kusaka-Kikushima, A., Tobiishi, M., Kawabata, K., Sayo, T., Sakai, S., Sugiyama, Y., Enomoto, H., Okada, Y., et al. (2013). KIAA1199, a deafness gene of unknown function, is a new hyaluronan binding protein involved in hyaluronan depolymerization. Proc. Natl. Acad. Sci. *110* , 5612–5617.

Zhang, F., Guo, X., Zhang, Y., Wen, Y., Wang, W., Wang, S., Yang, T., Shen, H., Chen, X., Tian, Q., et al. (2014). Genome-wide copy number variation study and gene expression analysis identify ABI3BP as a susceptibility gene for Kashin–Beck disease. Hum. Genet. *133*, 793–799.

Zhang, J., Haider, S., Baran, J., Cros, A., Guberman, J.M., Hsu, J., Liang, Y., Yao, L., and Kasprzyk, A. (2011). BioMart: a data federation framework for large collaborative projects. Database (Oxford) *2011*, bar038.

Zhang, K., Corsa, C.A., Ponik, S.M., Prior, J.L., Piwnica-Worms, D., Eliceiri, K.W., Keely, P.J., and Longmore, G.D. (2013). The collagen receptor discoidin domain receptor 2 stabilizes SNAIL1 to facilitate breast cancer metastasis. Nat. Cell Biol. *15*, 677–687.

Figure 1. Evaluation of Common MSC markers as transcriptional classifiers.

A) PLS-DA scatter plot of MSC (circles) and non-MSC cell types (triangles). Red symbols indicate cells which are incorrectly classified by the PLS-DA matrix. The matrix components consisted of 32 commonly used MSC markers.

B) Box and Whisker plots showing average expression of common MSC markers that are significantly differentially expressed (t-test, $P>10^{-6}$) between MSC (n=125) and non-MSC (n=510) cell types. See also Figure S1 and Table S1.

Figure 2. An improved *in silico* MSC signature.

A) Workflow summarizing the modified implementation of the sPLS-DA to integrate and evaluate cross-platform studies for derivation of a stable classifier;

B) Evaluation of the stability of each gene across four components, where frequency of selection over 200 subsamplings (Y-axis) is shown per gene (ENSEMBL ID, X-axis). Labels are provided for the 16 genes contributing to the signature across 4 components. Component 1 (green), Component 2 (Blue), Component 3 (Brown), Component 4 (Black). Small text gene symbols indicate a selection of previously identified MSC markers that were excluded for poor stability.

C) Evaluation of the contribution of each gene to the informativeness of its component. Each dot is a gene set, ordered along the x-axis by decreasing stability (frequency of selection). The y-axis represents the $-\log_{10}$(P-value) of a one tailed t-test indicating the improvement in classification accuracy across 4 components.

D) PLS-DA scatter plot showing sample clustering and classification accuracy of the training set (635 samples) in two components (Component 1 X axis, Component 2 Y-

686 axis). MSC samples are shown as circles, non-MSC as triangles, and misclassified
687 samples are coloured red.
688 E) Identifying the scores that classify an MSC or non-MSC. Distribution of the Rohart
689 MSC Score (X-axis) and the distribution density (Y-axis) for samples in the MSC
690 (n=115) or non-MSC (n=510) classes. Arrows indicate the scores that 99% of each
691 class fall into. The overlap indicates the region of uncertainty, where a classification is
692 given as 'unknown'.
693 F) A summary of the 16-gene MSC signature colour coded to the component (as
694 described in 1B). Gene ID is given as HUGO symbol and ENSEMBL gene ID; C is
695 component; P is probability of selection (indicating stability); R is correlation of gene to
696 component (as per 1D); L is predicted subcellular location of Intracellular (I), Nucleus
697 (N), Extracellular matrix (ECM), Secreted (S), Membrane (M) and U is unknown. See
698 also Supplementary Figure S2 and Supplemental Tables S2, S3.
699
700 **Figure 3**: **The MSC signature forms part of a network of extracellular proteins and**
701 **discriminates between differentiating or related adult stem cell types.**
702 A) An extended protein-protein network diagram of the Rohart MSC signature genes
703 demonstrating a role for VCAM1 and PDGFRB as part of a functionally interconnected
704 set of glycoproteins, integrins, growth factors and extracellular matrix proteins. Green
705 nodes are seed network members from component 1 genes, white nodes are inferred
706 network members, and edges are protein-protein interactions.
707 B) Box and Whisker plots showing average expression of the genes making up the
708 MSC signature component 1 genes in MSC (n=115) and non-MSC (n=510).

C) Classification of bone marrow MSC over a time course of differentiation to cartilage; y-axis gives the Rohart score, x-axis orders the samples from each experimental series. Three differentiation series from three donors are shown. The uncertainty region stands between the MSC and non-MSC prediction regions.

D) Classification of perivascular-derived stem cells from skeletal muscle mesangioblasts (HMAB), or iPSC-derived mesangioblasts (HIDEM) from donors with muscular dystrophy (MD) or healthy donors (WT). Error bars around each prediction score represent the CI boundaries. A sample is classified as 'unsure' (indicated in grey) if its prediction score or its CI overlapped the uncertainty region.

E) Classification of pericytes derived from three distinct datasets: from Left-Right neonatal foreskin (Antigen HD-1 dim or bright); placental pericytes; perivascular endometrial stem cells (CD146+/PDGRFB+). Stemformatics dataset identifiers provided for each experimental series. Error bars around each prediction score represent the CI boundaries.

F) Distribution of expression of common MSC/Pericyte markers. X-axis is Gene expression ranked by the YuGene cumulative proportion, Y-axis is the density distribution of MSC (orange plot, n=115) or pericytes (black plot, n= 16).

See also Supplemental Figure S3 and supplemental tables S4, S5 and S6.

Table 1(on next page)

Table 1

Table 1: MSC Signature improves the classification accuracy of MSC compared to a panel of 32 commonly used MSC markers. Column 1 provides the comparison of the classification accuracy of the 635 training samples using (Column 2) the 3 markers used as the minimal immunophenotype of the MSC training samples. (Column 3) a panel of 32 commonly used immune-markers in the MSC literature; (Column 4) using the unrefined sPLS-DA output; or (Column 5) with our final signature of 16 genes. Performance of each gene group was assessed using 200 random subsamplings of the training set. The internal classification error rate was calculated from a PLS-DA with 2 components (known immune-markers), or was an output of our statistical model with genes selected in an unbiased manner (cf Figure 1A).

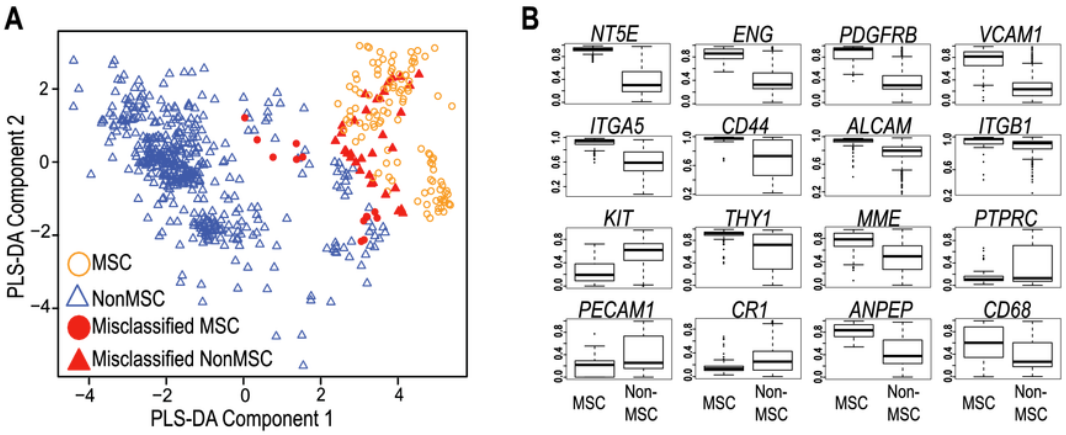
Table 1: MSC Signature improves the classification accuracy of MSC compared to a panel of 32 commonly used MSC markers. Column 1 provides the comparison of the classification accuracy of the 635 training samples using (Column 2) the 3 markers used as the minimal immunophenotype of the MSC training samples. (Column 3) a panel of 32 commonly used immune-markers in the MSC literature; (Column 4) using the unrefined sPLS-DA output; or (Column 5) with our final signature of 16 genes. Performance of each gene group was assessed using 200 random subsamplings of the training set. The internal classification error rate was calculated from a PLS-DA with 2 components (known immune-markers), or was an output of our statistical model with genes selected in an unbiased manner (cf Figure 1A).

	CD45, CD73, CD105	**32 common MSC markers**	**sPLS-DA prior to stable gene selection**	**The 16-gene MSC signature**
Overall accuracy (% of 635 samples)	87.86	92.33	97.71	97.85
MSC misclassified (% of 125 samples)	14.40	11.10	3.04	4.31
Non-MSC misclassified (% of 510 samples)	11.60	6.82	2.11	1.61

1

Figure 1.Evaluation of Common MSC markers as transcriptional classifiers.

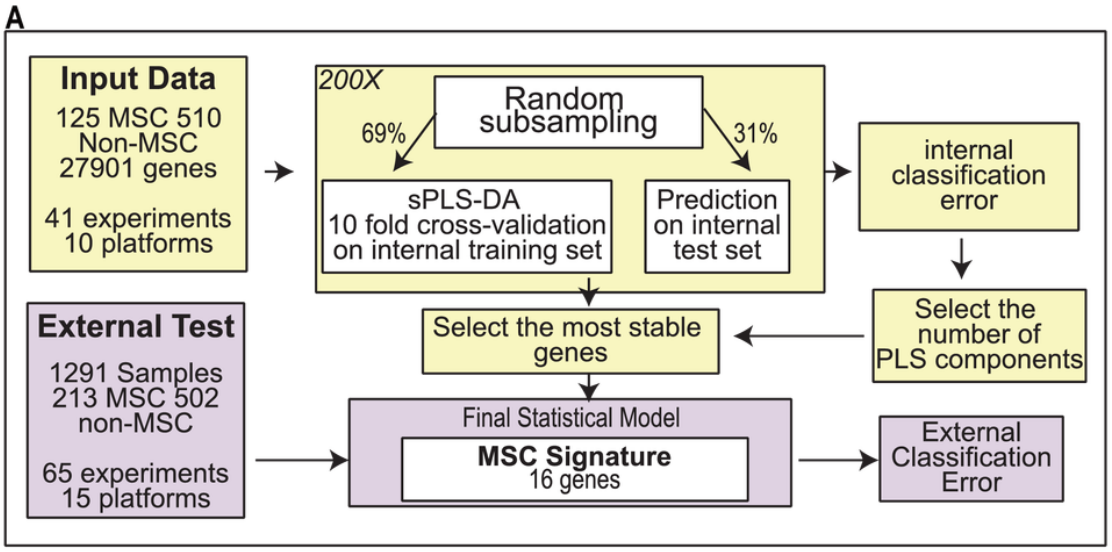
A) PLS-DA scatter plot of MSC (circles) and non-MSC cell types (triangles). Red symbols indicate cells which are incorrectly classified by the PLS-DA matrix. The matrix components consisted of 32 commonly used MSC markers. B) Box and Whisker plots showing average expression of common MSC markers that are significantly differentially expressed (t-test, $P>10^{-6}$) between MSC (n=125) and non-MSC (n=510) cell types. See also Figure S1 and Table S1.

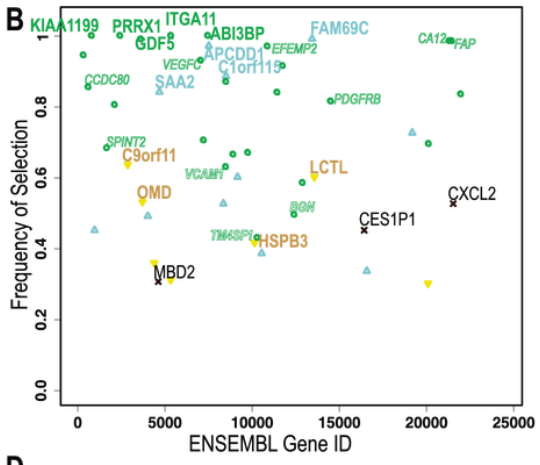

A
PLS-DA Component 2
PLS-DA Component 1
MSC
NonMSC
Misclassified MSC
Misclassified NonMSC
B
NT5E
ENG
PDGFRB
VCAM1
ITGA5
CD44
ALCAM
ITGB1
KIT
THY1
MME
PTPRC
PECAM1
CR1
ANPEP
CD68
MSC
Non-MSC

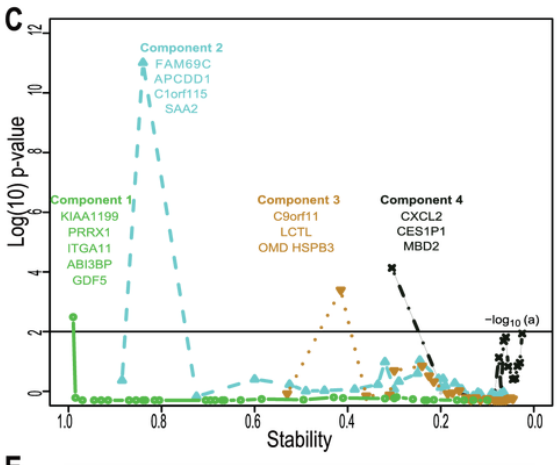
2

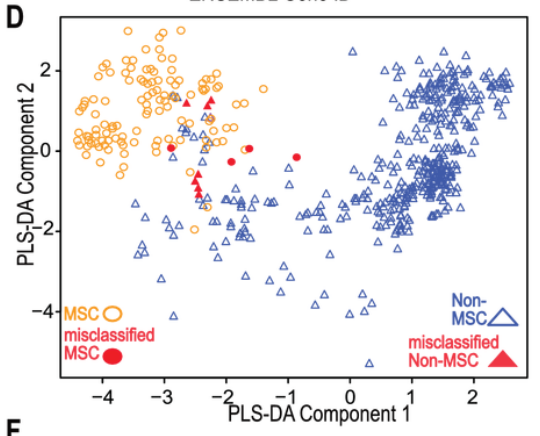
Figure 2.An improved *in silico* MSC signature.

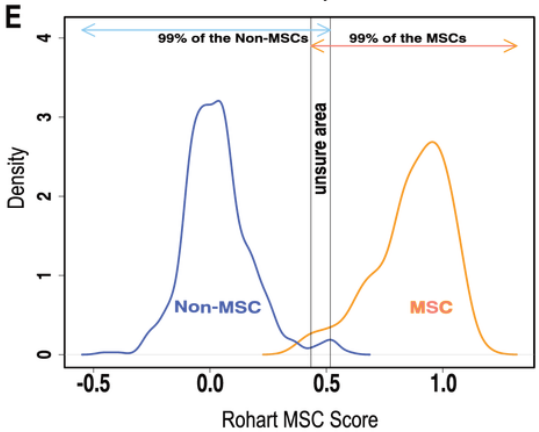
A) Workflow summarizing the modified implementation of thesPLS-DA to integrate and evaluate cross-platform studies for derivation of a stable classifier; B) Evaluation of the stability of each gene across four components, where frequency of selection over 200 subsamplings (Y-axis) is shown per gene (ENSEMBL ID, X-axis). Labels are provided for the 16 genes contributing to the signature across 4 components. Component 1 (green), Component 2 (Blue), Component 3 (Brown), Component 4 (Black). Small text gene symbols indicate a selection of previously identified MSC markers that were excluded for poor stability. C) Evaluation of the contribution of each gene to the informativeness of its component. Each dot is a gene set, ordered along the x-axis by decreasing stability (frequency of selection). The y-axis represents the $-\log_{10}$(P-value) of a one tailed t-test indicating the improvement in classification accuracy across 4 components. D) PLS-DA scatter plot showing sample clustering and classification accuracy of the training set (635 samples) in two components (Component 1 X axis, Component 2 Y-axis). MSC samples are shown as circles, non-MSC as triangles, and misclassified samples are coloured red. E) Identifying the scores that classify an MSC or non-MSC. Distribution of the Rohart MSC Score (X-axis) and the distribution density (Y-axis) for samples in the MSC (n=115) or non-MSC (n=510) classes. Arrows indicate the scores that 99% of each class fall into. The overlap indicates the region of uncertainty, where a classification is given as ‘unknown’. F) A summary of the 16-gene MSC signature colour coded to the component (as described in 1B). Gene ID is given as HUGO symbol and ENSEMBL gene ID; C is component; P is probability of selection (indicating stability); R is correlation of gene to component (as per 1D); L is predicted subcellular location of Intracellular (I), Nucleus (N), Extracellular matrix (ECM), Secreted (S), Membrane (M) and U is unknown. See also Supplementary Figure S2 and Supplemental Tables S2, S3.











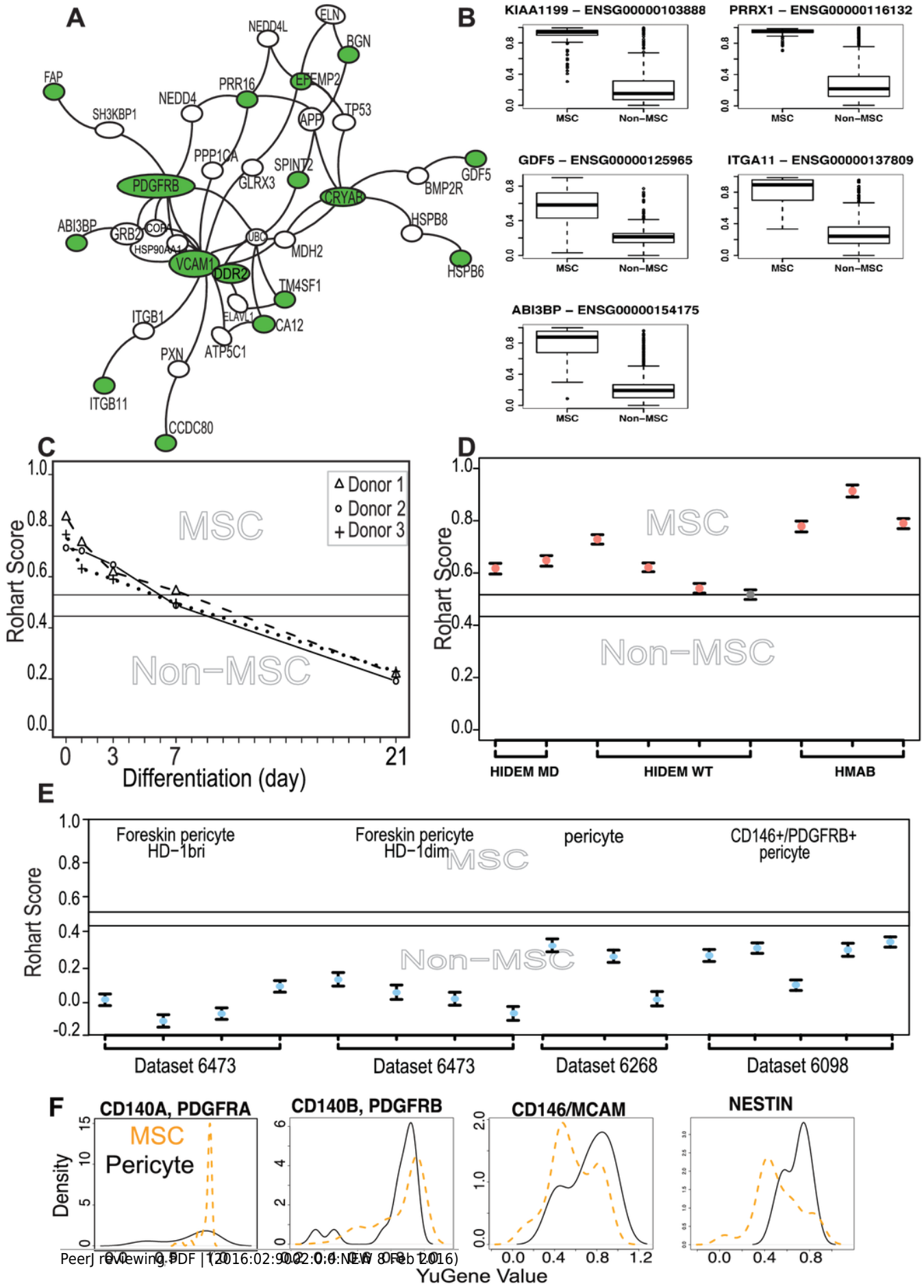
F

ENSEMBL Gene ID	C	P	R	L	SYMBOL	Description
ENSG00000134046	1	1	-0.93	I, ECM	KIAA1199	Cell migration inducing protein, CEMIP
ENSG00000116132	1	1	-0.94	N	PRRX1	Paired related homeobox 1
ENSG00000137809	1	1	-0.83	M	ITGA11	Integrin alpha 11, beta 1
ENSG00000154175	1	1	-0.91	ECM	ABI3BP	ABI family member 3 (NESH) Binding Protein, TARSH
ENSG00000125965	1	0.99	-0.92	S	GDF5	Growth differentiation factor 5, BMP14
ENSG00000187773	2	0.99	-0.55	I	FAM69C	Cysteine-rich type II transmembrane protein
ENSG00000154856	2	0.96	-0.70	M	APCDD1	Adenomatosis polyposis coli down regulated 1
ENSG00000162817	2	0.89	-0.74	U (M)	C1orf115	Chromosome 1 uncharacterised open reading frame 115
ENSG00000134339	2	0.82	-0.80	S	SAA2	Serum amyloid A2
ENSG00000120160	3	0.66	0.63	I	C9orf11	Sperm acrosome associated, Equatorin
ENSG00000188501	3	0.61	-0.68	I	LCTL	Lactase-like 1
ENSG00000127083	3	0.57	-0.42	ECM	OMD	Osteomodulin
ENSG00000169271	3	0.41	0.58	N	HSPB3	Heat shock 27kDa protein 3
ENSG00000081041	4	0.53	0.54	S	CXCL2	Chemokine (C-X-C motif) ligand 2
ENSG00000204382	4	0.42	0.62	N	CES1P1	Pseudogene, Carboxylesterase 1
ENSG00000134046	4	0.31	0.74	N	MBD2	Methyl-CpG binding protein 2

3

Figure 3: The MSC signature forms part of a network of extracellular proteins and discriminates between differentiating or related adult stem cell types.

A) An extended protein-protein network diagram of the Rohart MSC signature genes demonstrating a role for VCAM1 and PDGFRB as part of a functionally interconnected set of glycoproteins, integrins, growth factors and extracellular matrix proteins.Green nodes are seed network members from component 1 genes, white nodes are inferred network members, and edges are protein-protein interactions. B) Box and Whisker plots showing average expression of the genes making up the MSC signature component 1 genes in MSC (n=115) and non-MSC (n=510). C) Classification of bone marrow MSC over a time course of differentiation to cartilage; y-axis gives the Rohart score, x-axis orders the samples from each experimental series. Three differentiation series from three donors are shown. The uncertainty region stands between the MSC and non-MSC prediction regions. D) Classification of perivascular-derived stem cells from skeletal muscle mesangioblasts (HMAB), or iPSC-derived mesangioblasts (HIDEM) from donors with muscular dystrophy (MD) or healthy donors (WT). Error bars around each prediction score represent the CI boundaries. A sample is classified as ‘unsure’ (indicated in grey) if its prediction score or its CI overlapped the uncertainty region. E) Classification of pericytes derived from three distinct datasets: from Left-Right neonatal foreskin (Antigen HD-1 dim or bright); placental pericytes; perivascular endometrial stem cells (CD146+/PDGRFB+). Stemformatics dataset identifiers provided for each experimental series. Error bars around each prediction score represent the CI boundaries. F) Distribution of expression of common MSC/Pericyte markers. X-axis is Gene expression ranked by the YuGene cumulative proportion, Y-axis is the density distribution of MSC (orange plot, n=115) or pericytes (black plot, n= 16). See also Supplemental Figure S3 and supplemental tables S4, S5 and S6.


A
NEDD4L
ELN
BGN
FAP
PRR16
EFEMP2
NEDD4
SH3KBP1
TP53
APP
PPP1CA
SPINT2
GDF5
PDGFRB
GLRX3
BMP2R
CRYAB
ABI3BP
GRB2
COPA
HSPB8
UBC
HSP90AA1
MDH2
VCAM1
DDR2
HSPB6
TM4SF1
ITGB1
ELAVL1
CA12
ATP5C1
PXN
ITGB11
CCDC80
B
KIAA1199 – ENSG00000103888
PRRX1 – ENSG00000116132
GDF5 – ENSG00000125965
ITGA11 – ENSG00000137809
ABI3BP – ENSG00000154175
MSC
Non−MSC
C
Donor 1
Donor 2
Donor 3
MSC
Non−MSC
Rohart Score
Differentiation (day)
D
MSC
Non−MSC
Rohart Score
HIDEM MD
HIDEM WT
HMAB
E
Foreskin pericyte HD−1bri
Foreskin pericyte HD−1dim
pericyte
CD146+/PDGFRB+ pericyte
MSC
Non−MSC
Rohart Score
Dataset 6473
Dataset 6473
Dataset 6268
Dataset 6098
F
CD140A, PDGFRA
CD140B, PDGFRB
CD146/MCAM
NESTIN
MSC
Pericyte
Density
YuGene Value