Chapter 2 - Materials and Methods

Experimental procedures

Experiments associated with the data presented in this thesis were performed by Roser Vento-Tormo with help from Margherita Turco, Rachel Botting, Jongeun Park, and Rebecca Payne, and are described in Appendix 1.

Data analysis

I performed all analyses described in the following sections with the exception of the work described under “Single-cell RNA-sequencing data analysis,” which was performed by Mirjana Efremova.

Single-cell RNA-sequencing data analysis

Droplet-based scRNA-seq data was aligned and quantified using the Cell Ranger Single-Cell Software Suite (v.2.0)[110] against the GRCh38 human reference genome. Cells with fewer than 500 detected genes or more than 20% mitochondrial gene content were removed. Genes expressed in fewer than 3 cells were also removed. SmartSeq2 sequencing data was aligned with HISAT2[112] using the same genome reference and annotation as the droplet-based data. Gene-specific read counts were calculated using HTSeq-count[113]. Cells with fewer than 1,000 detected genes or more than 20% mitochondrial gene content were removed. Genes expressed in fewer than 3 cells were also removed. Downstream analyses such as gene expression log-normalization, k-nearest neighbor graph clustering, differential expression analyses (Wilcoxon rank-sum test), and visualization using the t-SNE algorithm[114] were performed using the R package Seurat v.2.1.0[115]. t-SNE analyses were performed using a perplexity of 30. Clusters were annotated based on expression of canonical cell type markers listed in Appendix 2. We further removed cells we did not gate for (most likely maternal blood B cells and fetal brain tissue), clusters for which the top markers were genes associated with dissociation-induced effects[116], or mitochondrial genes, and a fibroblast cluster with high expression of hemoglobin genes due to background contamination of cell-free RNA.

Maternal-fetal single-cell genotyping

Whole-genome sequencing alignment and variant calling:

Maternal and fetal whole-genome sequencing data were mapped to the GRCh37.p13 reference genome using BWA-MEM v.0.7.15[117]. The SAMtools[118] fixmate utility v.1.5 was used to update read pairing information and mate-related flags. Reads near known indels from the Mills[119] and 1000G[120] gold standard reference set for GRCh37/hg19 were locally realigned using GATK IndelRealigner v.3.7[121] (*-model* KNOWNS_ONLY *-LOD* 0.4). Base calling assessment and base quality scores were adjusted with GATK BaseRecalibrator and PrintReads v.3.7[121]. PCR duplicates were identified and removed using Picard MarkDuplicates v.2.14.1[121,122]. Finally, bcftools mpileup and call v.1.6[123] were used to produce genotype likelihoods and output called variants at all known biallelic SNP sites overlapping protein-coding genes, compiled from NCBI dbSNP build 138 (GRCh37/hg19)[124]. For each sample, variants called with phred-scale quality score (QUAL) ≥ 200, at least 20 supporting reads (DP ≥ 20), and mapping quality (MQ) ≥ 60 were retained as high-quality variants.

Inferring maternal/fetal genetic origin of single cells from droplet-based scRNA-seq using whole-genome sequencing variant calls:

To match the processing of the whole-genome sequencing datasets, droplet-based sequencing data from decidua and placenta samples were realigned and quantified against the GRCh37 human reference genome using the Cell Ranger Single-Cell Software Suite (v.2.0)[110]. The fetal or maternal origin of each barcoded cell was then determined using the tool demuxlet[125]. Briefly, demuxlet can be used to deconvolute droplet-based scRNA-seq experiments in which cells are pooled from multiple, genetically distinct individuals. Given a set of genotypes corresponding to these individuals, demuxlet infers the most likely genetic identity of each droplet by evaluating scRNA-seq reads from the droplet which overlap known SNPs. Demuxlet inferred the identities of cells in this study by analyzing each Cell Ranger-aligned BAM file from decidua or placenta in conjunction with a VCF containing the high-quality variant calls from the corresponding WGS of maternal and fetal DNA (*--field* GT). Each droplet was assigned to be maternal, fetal, or unknown in origin (ambiguous or potential doublet), and these identities were then linked with the transcriptome-based cell clustering data to confirm the maternal and fetal identity of each annotated cell type.

Inferring maternal/fetal genetic origin of single cells from droplet-based scRNA-seq read data alone:

10x Chromium droplet-based sequencing data from decidua and placenta samples were realigned and quantified against the GRCh37 human reference genome using STAR v.2.2.1[126] with the following parameters: *--alignSJoverhangMin* 8, *--alignSJDBoverhangMin* 1, *--alignIntronMin* 20 *--alignIntronMax*

1000000, *--alignMatesGapMax* 1000000, *--sjdbScore* 2, *--outFilterType* BySJout, *--outFilterMultimapNmax* 20, *--outFilterMismatchNmax* 999, *--outFilterMismatchNoverLmax* 0.04, *--outFilterScoreMinOverLread* 0.33, *--outFilterMatchNminOverLread* 0.33, *--outSAMstrandField* intronMotif, *--outFilterIntronMotifs* RemoveNoncanonical, *--outSAMattributes* NH HI NM MD AS XS, *--outSAMunmapped* Within, *--twopassMode* Basic.

Reads near known indels from the Mills[119] and 1000G[120] gold standard reference set for GRCh37/hg19 were locally realigned using GATK IndelRealigner v.3.7[121] (*-model* KNOWNS_ONLY *-LOD* 0.4). Base calling assessment and base quality scores were adjusted with GATK BaseRecalibrator and PrintReads v.3.7[121]. Next, reads in each sample BAM file were split by Chromium cellular barcode to produce a separate BAM file for each single cell from the sample. For each single-cell BAM file, PCR duplicates were identified and removed using Picard MarkDuplicates v.2.14.1[121,122]. Finally, GATK HaplotypeCaller v.3.7[121] was used to produce genotype likelihoods and output called variants for each cell based on reads containing known biallelic SNP sites from NCBI dbSNP build 138 (GRCh37/hg19)[124] overlapping the top 1000 genes most highly expressed in placental and endometrium RNA-seq data deposited in the Human Protein Atlas[127,128].

The vcf files from decidual and placental single cells were merged and the R/Bioconductor package vcfR v.1.6.0[129] was used to import the merged vcf into R as a sparse matrix. We performed filtering on the SNPs so that only SNPs called in more than one cell and with non-zero variance were retained for downstream analysis. Next, using the R/Bioconductor package pcaMethods v.1.68.0[130], we performed a probabilistic PCA (“ppca”) on the SNP data with unit variance scaling (“uv”) on two principal components and visualized the resulting projections using ggplot2, colored by tissue of origin or previously inferred cell identities.

Decidua bulk RNA-sequencing processing and heat shock protein expression analysis

Reads from five decidua bulk RNA-seq samples were mapped and quantified against the GRCh38 (release 88) human reference transcriptome using the lightweight-alignment (SMEM-based) mode in Salmon v.0.8.1[131]. The R/Bioconductor package tximport v.1.4.0[132] was used to aggregate the transcript-level abundances into gene-level expression estimates (TPM) with Ensembl gene IDs (GRCh38) as the primary identifier. Ensembl IDs were then mapped to HGNC gene symbols using R/Bioconductor package biomaRt v.2.32.1[133]. We specifically examined the expression of heat shock protein-related genes *HSPA6, DNAJB1, HSPH1, DNAJA4, HSP90AA1, HSPA1A, HSPA1B, HSPD1, DNAJA1, HSPA8,* and *HSPB1* in the bulk RNA-seq datasets. These were among the most highly upregulated genes in the T3 T cell cluster identified from analysis of the decidua plate-based scRNA-seq data. *HSPB1, HSPA1A,* and *HSPA1B* were also among the genes previously found to be induced by single-cell dissociation protocols[109].

FACS and SmartSeq2 data analysis

FACS data were gated and compensated in FlowJo and exported as FCS files. Gating identities coupled with plate locations from index sorting were then imported into R using the Bioconductor package flowCore v.1.42.3[134] and linked to the metadata for each corresponding cell generated from Seurat analyses of decidua and peripheral blood SmartSeq2 data (as performed in "Single-cell RNA-sequencing data analysis"). This facilitated superimposition of the gated identities of cells onto the t-SNE projections defined by single-cell transcriptomes. Differential expression analysis between the dMP1 and dMP2 subsets was performed using the Seurat *FindMarkers* function (Wilcoxon rank-sum, among genes expressed in at least 10% of cells).

CyTOF data analysis

Populations of interest were manually gated in FlowJo and exported as FCS files. Subsequent analyses were conducted with the Bioconductor package cytofkit v.1.8.4[135]. First, signal intensities for each marker were transformed using the negative value pruned inverse hyperbolic sine transformation (cytofAsinh). To obtain two-dimensional visualizations of the CyTOF data, the t-SNE algorithm[114] was applied to 10,000 randomly selected cells from each dataset and plotted using the R package ggplot2. Marker expression was visualized on the t-SNE plots, with maximum intensity designated as marker expression intensities in the 99th percentile or higher.

Gene Ontology and Reactome term enrichment analysis

Genes significantly upregulated ($\log_2$(fold-change) ≥ 0.5, adjusted $p < 0.05$, Wilcoxon rank-sum) in each of the mononuclear phagocyte populations relative to other cell types at the maternal-fetal interface were functionally annotated using gene ontology (GO)[136] and Reactome[137] pathway terms. We used the R/Bioconductor package gProfileR v.0.6.4[138] to map gene lists ordered by decreasing $\log_2$(fold-change) (*ordered_query* = T) to biological pathway (BP) and molecular function (MF) GO terms and Reactome pathway annotations. Through gProfileR we then performed statistical enrichment analysis to identify and output overrepresented terms with strong hierarchical filtering (*hier_filtering* = “strong”). All *p*-values were corrected for multiple testing using the gProfileR gSCS algorithm[138].

Intersection of M1/M2 gene signatures with maternal mononuclear phagocyte subset markers

We obtained a list of canonical M1 macrophage and M2 macrophage marker genes[139] and used biomaRt v.2.32.1[133] to map the gene symbols to corresponding Ensembl gene IDs. We then intersected this gene list with our lists of upregulated genes (adjusted $p < 0.05$, Wilcoxon rank-sum) from each of the mononuclear phagocyte populations relative to other decidual and placental cell types. Heatmaps showing single cell-level expression of the M1/M2 genes were plotted using the *DoHeatmap* function in Seurat, with genes presented in decreasing $\log_2$(fold-change), grouped by the mononuclear phagocyte population in which they were significantly upregulated. If a gene was upregulated in multiple mononuclear phagocyte subsets, it was grouped with the subset in which it exhibited the highest $\log_2$(fold-change) relative to other cell types.

Identification of placenta- and endometrium-specific genes and intersection with maternal resident immune cell population markers

Tissue-level RNA-seq data were downloaded from the Human Protein Atlas (www.proteinatlas.org)[127,128]. We obtained gene expression data (in tpm) for 37 human tissues sourced from 122 individuals; the associated experiment accession for this dataset in the ArrayExpress database is E-MTAB-2836.

To identify genes specifically enriched in the placenta and endometrium relative to other tissues, we employed the tissue specificity metric Tau[140], which was determined to be among the most robust methods for determining tissue-specific gene expression patterns in a recent comparative analysis[141] with other approaches, including expression enrichment (EE)[142], Hg (Shannon entropy)[143], tissue-specificity index (TSI)[144], and z-score[145]. For each gene, we calculated Tau using the following formula:

$$\tau = \frac{\sum_{i=1}^{n}\left(1-\widehat{x_i}\right)}{n-1}; \;\; \widehat{x_i} = \frac{x_i}{\max\limits_{1\leq i\leq n}(x_i)}.$$

where x_i is the expression of gene in tissue *i* and *n* is the total number of tissues. Genes with Tau ≥ 0.8[141] and tissue of highest level of expression being placenta or endometrium were determined to be significantly enriched in these tissues. A list of the 410 placenta- and endometrium-enriched genes is located in Appendix 5.

We then intersected this gene list with our lists of upregulated genes (adjusted $p < 0.05$, Wilcoxon rank-sum) from each of the maternal resident immune cell populations (T cells, NK cells, mononuclear phagocytes) relative to other decidual and placental cell types. Heatmaps showing single cell-level expression of the placenta- and endometrium-enriched genes were plotted using the *DoHeatmap* function in Seurat v.2.1.0, with genes presented in decreasing $\log_2$(fold-change), grouped by the maternal resident immune cell type in which they were significantly upregulated. If a gene was determined to be upregulated in maternal resident immune cell subsets, it was grouped with the population in which it exhibited the highest $\log_2$(fold-change) relative to other cell types.

Curation of genes associated with fertility or complications of pregnancy and analysis of cell-type specific expression at the maternal-fetal interface

Genes associated with abnormal birth weight or fetal growth, endometriosis/ovarian disease, gestational trophoblastic disorder/hydatidiform mole, preeclampsia, age of menopause or menstrual onset, preterm birth, recurrent miscarriage, placental abruption, and placenta accreta were curated from studies deposited in the NHGRI/EBI GWAS Catalog[146], OMIM database[147], and from literature searches. All genes selected from literature were linked with increased mutation or with alterations in expression or epigenetic regulation in studies of a particular condition in humans or human tissues. A full table of compiled genes, along with their associated conditions and their literature or database sources, is provided in Appendix 6.

We first intersected our curated gene list with our lists of upregulated genes (adjusted $p < 0.05$, Wilcoxon rank-sum) from each of the cell populations at the maternal-fetal interface relative to other decidual and placental cell types. Heatmaps showing cell type-averaged expression of the disease- or fertility-associated genes were plotted using the *heatmap* function in R, with genes presented in decreasing $\log_2$(fold-change), grouped with the cell type in which they were significantly upregulated. If a gene was determined to be upregulated in multiple cell types, it was grouped with the cell type in which it exhibited the highest $\log_2$(fold-change) relative to other cell populations. For our maternal- and fetal-specific analyses, we first determined which genes were significantly upregulated (adjusted $p < 0.05$) in each maternal cell type relative to other maternal cell types, or in each fetal cell type relative to other fetal cell types, using the Seurat *FindMarkers* function (Wilcoxon rank-sum, among genes expressed in at least 10% of cells). We then intersected our curated gene lists with these maternal- and fetal-specific upregulated gene lists and plotted cell type-averaged gene expression using the *heatmap* function in R as previously described.