DESCRIPTION OF TABLES S1-S18

https://doi.org/10.1093/nar/gkad339

Table S1:

Sheet Name	Summary of Contents
BBBsamples	Information on samples obtained from the Brisbane Breast Bank (BBB).
FACSmethods	Antibodies and immuno-staining conditions for FACS experiments.

Table S2:

Sheet Name	Summary of Contents
SampleDescription	Contains the identifiers for all sequenced samples, including the original Novogene file name, short name for internal identification and descriptive name containing cell type and patient origin.
ReadNumbers	Overal number of reads sequenced per sample and final average of all samples.
MetricsFastQC_R1	Main metrics obtained by FastQC (version 0.11.5) for reads in R1 files, before any processing or filtering (raw reads).
MetricsFastQC_R2	Main metrics obtained by FastQC (version 0.11.5) for reads in R2 files, before any processing or filtering (raw reads).

Table S3:

Sheet Name	Summary of Contents
ReadNumbers	Overal number of reads sequenced per sample and final average of all samples.
MetricsFastQC_R1	Main metrics obtained by FastQC (version 0.11.5) for reads in R1 files, after trimming.
MetricsFastQC_R2	Main metrics obtained by FastQC (version 0.11.5) for reads in R2 files, after trimming.

Table S4:

Sheet Name	Summary of Contents
Strandness	Percentage of reads identified as RSeQC as being of the type "1+-,1-+,2++,2--".
CorrectedReads	Percentage of reads for which Rcorrector attempted correction.
UncorrectableReads	Percentage of reads FurPe deemed uncorrectable after Rcorrector attempts.
TrimmingSummary	Percentage of reads surviving trimming either in a pair or individually. Percentages are presented for reads before and after correction efforts, for comparison.
RibosomalReads	Percentage of reads coming from ribosomal RNAs (as annotated in the SilvaDB) in each sample in reads before and after correction, for comparison.
MappedReads	Overall alignment rate between reads and the human genome (GENCODE hg38).

Table S5:

Sheet Name	Summary of Contents
AssemblyMetrics	Main assembly metrics calculated by TrinStats for the raw assembly and the TransRate filtered assembly.
TransRateMetrics	Trinity assembly statistics before and after TransRate filtering for optimisation.
BUSCOmetrics	Results for BUSCO using both the Eukaryote and the Mammalian sets of orthologs.
Comparison with Holzer and Marz	Comparisons between our in-house assembly and data from Hölzer and Marz, 2019.

Table S6:

Sheet Name	Summary of Contents
CodingPotential	Summary of all coding potential calculators used in this work.
FEELnc	Summary of FEELnc results for genes with supported lack of coding potential (only 'best' targets are shown).
ClasseslncRNAs	Number of lncRNAs per class (defined by their genomic context) used to create the included pie chart.

Table S7:

Sheet Name	Summary of Contents
lncRNAs_with_LongReadSupport_A	Summary of BLAT alignment between NB-lncRNAs and publicly available long-reads (after applying coverage filter).
lncRNAs_with_LongReadSupport_B	Summary of BLAT alignment between NB-lncRNAs and long-reads generated in-house (after applying coverage filter).
NumberOfLongReads_per_lncRNA	Number of long-reads that align to each lncRNA (after applying coverage filter).
lncRNAs_with_TSS_A	Summary of Bedtools intersect between lncRNAs and TSS peaks from publicly available RAMPAGE data (after filtering for TSSs located <500bp upstream or <50bp internal to transcript start).
lncRNAs_with_TSS_B	Summary of Bedtools intersect between lncRNAs and TSS peaks from in-house RAMPAGE experiments (after filtering for TSSs located <500bp upstream or <50bp internal to transcript start).
RT-PCR_primers	Primers designed for RT-PCT of NB-lncRNAs presented at Fig. S1b

Table S8:

SheetName	Summary of Contents
NBdb	Normal Breast DataBase of genes from the literature previously assigned to each cell subpopulation.
NBdb_Statistics	Statistics of the database, including the number of genes per cell subpopulation and genes per reference.
NBdb_GeneAnnotation	Genomic annotation of all genes in the database.
InHouse_lncRNAdb	Database of unannotated lncRNAs compiled from multiple sources.
Telomere_geneDB	Genes implicated in telomere maintenance listed in GSEA (Reactome and Biocarta gene sets).
DNArepair_geneDB	Genes implicated in DNA repair pathways listed in REPAIRtoire and/or MD Anderson databases.
EMT_geneDB	EMT-related genes from GSEA and/or EMTome.
ClaudinLow_geneDB	Genes upregulated in claudin-low samples, according with Prat, 2010.
HumanHousekeepingGenes	Human housekeeping genes, deposited at HK and/or HRT.
RibosomalRNA	Sequences for rRNA depletion (169 entries of rRNAs).

Table S9:

Sheet Name	Summary of Contents
GencodeAnnotated	List of all assembled lncRNAs annotated in hg38 human genome assembly from GENCODE.
SummaryOfGencodeAnnotated	Overall counts of genes from different types from GencodeAnnotated.
ncRNAdbAnnotation	List of NB-lncRNAs annotated in the in-house database of ncRNAs.
Lnc2CancerMatches	List of NB-lncRNAs annotated in the Lnc2Cancer database.
Lnc2CancerClasses	Classification of Lnc2CancerMatches NB-lncRNAs in Lnc2Cancer.
lncRNAfuncMatches	List of NB-lncRNAs annotated as genes featuring in lncRNAfunc.
MaTARs	Correspondence between MaTARs and NB-lncRNAs.

Table S10:

Sheet Name	Summary of Contents
349_elncRNAs_AndPairedENhancers	List of 349 lncRNAs that co-localize with known enhancer regions.
TargetsOf_elncRNAs	Pairwise correlation between elncRNAs and the correspondent annotated enhancer element.
NBmarkers_TargetedBy_elncRNAs	Normal breast marker genes targeted by the elncRNAs.
1968pancRNAs_AndPairedPromoters	List of 1968 lncRNAs that co-localize with known promoter regions.
TargetsOf_parcRNAs	Pairwise correlation between pancRNAs and the correspondent annotated promoters.
NBmarkers_TargetedBy_parcRNAs	Normal breast marker genes targeted by the parcRNAs.
825TALRs_AndPairedUTRs	List of 825 TALRs that co-localize with known UTRs.

TargetsOf_TALRs	Pairwise correlation between TALRs and the correspondent annotated UTRs.
NBmarkers_TargetedBy_TALRs	Normal breast marker genes targeted by the TALRs.

Table S11:

Sheet Name	Summary of Contents
ConsistentlyExpressed_in_C1	lncRNAs consistently expressed in the C1 population (above 1 TPM)
ConsistentlyExpressed_in_C2	lncRNAs consistently expressed in the C2 population (above 1 TPM)
ConsistentlyExpressed_in_C4	lncRNAs consistently expressed in the C4 population (above 1 TPM)
NBmarkersOfInterest	Population-specific NB-lncRNAs that are partners of known protein-coding markers
NBmarkers_TargetsOf_CE_in_C1	Normal breast marker genes targeted by the lncRNAs consistently expressed in the C1 population
NBmarkers_TargetsOf_CE_in_C2	Normal breast marker genes targeted by the lncRNAs consistently expressed in the C2 population
NBmarkers_TargetsOf_CE_in_C4	Normal breast marker genes targeted by the lncRNAs consistently expressed in the C4 population
NBmarkers_TargetsOf_UCE_in_C1	Normal breast marker genes targeted by the lncRNAs consistently expressed only in the C1 population
NBmarkers_TargetsOf_UCE_in_C2	Normal breast marker genes targeted by the lncRNAs consistently expressed only in the C2 population
NBmarkers_TargetsOf_UCE_in_C4	Normal breast marker genes targeted by the lncRNAs consistently expressed only in the C4 population

Table S12:

Sheet Name	Summary of Contents
FACS_CellLabels	Flow cytometry labels provided in Nguyen, 2018 for each cell.
ClusterAnnotation_FACS	Based only on information from the FACS labels, how we characterized each cluster and the corresponding confusion matrices.
PhysiologicalCharacteristics	Overlap between marker genes and genes in investigated physiological characteristics of basal and luminal cell types.
ClusterAnnotation_LabMarkers	Based on currently used laboratory clusters, how would each cluster be annotated.
BasalMarkers_Pal2021	List of basal markers from Pal et al., 2021
LumProgenitorMarkers_Pal2021	List of luminal progenitor markers from Pal et al., 2021
LumMatureMarkers_Pal2021	List of luminal mature markers from Pal et al., 2021
RefClusters_LiteratureMarkers	List of literature markers from the in-house database and which cluster has each gene as Serurat marker.
ClusterCorrespondence	How cells from A-clusters were divided into L-clusters.

Table S13:

Sheet Name	Summary of Contents
SeuratMarkers_Aclusters	List of markers assigned by Seurat for each cluster obtained based on GENCODE-annotated gene expression (A-clusters).
SeuratMarkers_Lclusters	List of markers assigned by Seurat for each cluster obtained based on NB-lncRNA expression (L-clusters).
SeuratMarkers_Mclusters	List of markers assigned by Seurat for each cluster obtained based on merged GENCODE-annotated gene expression and NB-lncRNA expression (M-clusters).
SeuratMarkers_Oclusters	List of markers assigned by Seurat for each cluster obtained based on the expression of GENCODE-annotated genes that are neither confirmed protien-coding or lncRNAs (O-clusters).
SeuratMarkers_AnnotatedLncRNAs	List of markers assigned by Seurat for each cluster obtained based on GENCODE-annotated lncRNAs.

Table S14:

Sheet Name	Summary of Contents
MarkersOfInterest_NBdb_FEELnc	Seurat-assigned NB-lncRNA markers with FEELnc-predicted targets in the database of normal breast marker genes.
MarkersOfInterest_BroadCellType	Seurat-assigned NB-lncRNA markers expressed in all subpopulations of each cell type.
MarkersOfInterest_UniqToCluster	Seurat-assigned NB-lncRNA markers that have expression confined to each cluster.

Table S15:

Sheet Name	Summary of Contents
WidespreadExpression_NBlncRNAs	List of NB-lncRNAs expressed in more than 1/3 of the cells and their FEELnc-assigned targets.
WidespreadExpression_PCGs	List of GENCODE-annotated genes expressed in more than 1/3 of the cells and their gene names.
HumanHousekeepingGenes	Human housekeeping genes, deposited at HK and/or HRT.
Overlap_WE_NBlncRNA_HK_GeneDB	Overlap between the list of NB-lncRNAs of widespread expression and housekeeping genes (Sheet 4). The overlaps are shown for genes co-expressed with NB-lncRNAs (left) and for predicted targets of the NB-lncRNAs (right).

Table S16:

Sheet Name	Summary of Contents
Brain_SeuratMarkers_Aclusters	List of markers assigned by Seurat for each cluster obtained based on GENCODE-annotated gene expression (A-clusters).
Brain_SeuratMarkers_Lclusters	List of markers assigned by Seurat for each cluster obtained based on NB-lncRNAs expression (L-clusters).
CellTypes_PerCluster	Distribution of cell types (defined based on provided labels from original files) in each cluster.
KnownMarkers_PerCluster	Number of known markers (in Darmanis, 2015) in each cluster.
MarkersOfInterest	Seurat-assigned markers not necessarily mentioned in Darmanis, 2015 which we found to be of interest.

Table S17:

Sheet Name	Summary of Contents
SelectedStemCellMarkers	Contains the list of genes selected for root-state determination, based on experimental data.
ClusterSRvalues	SR values calculated by SCENT for each cell, averaged per cluster.

Table S18:

Sheet Name	Summary of Contents
PearsonCorrelation_Cluster_TCGA	For each L-cluster, the Pearson correlation coefficient and associated p-value for its correspondence with each TCGA subtype.
TCGAmarkers_NBlncRNAs	NB-lncRNAs defined in-house as markers of TCGA subtypes, for each subtype the top 300 most frequently assigned markers.
TCGAmarkers_GENCODEannotated	GENCODE-annotated genes defined in-house as markers of TCGA subtypes (top 300 most frequently assigned markers per subtype).
Correspondence_wWuetalMarkers	P-values of overlaps between in-house markers of TCGA subtypes and markers in Wu et al., 2021, for both GENCODE-annotated and NB-lncRNA markers.
Correspondence_wSeuratlMarkers	P-values of overlaps between in-house NB-lncRNA markers of TCGA subtypes and Seurat-assigned markers of L-clusters.
NBlncRNAs_MarkersOfInterest	List of MGFR-assigned NB-lncRNA markers used for subtype separation with connection to breast cancer or other features of interest. The top markers contributing with the first two principal components are bolded.