SummarizedExperiment - se

La classe SummarizedExperiment è utilizzata per memorizzare matrici rettangolari di risultati sperimentali, comunemente ottenuti da esperimenti di sequenziamento e microarray. SummarizedExperiment può gestire simultaneamente diversi risultati sperimentali o saggi, purché abbiano le stesse dimensioni.

Ogni oggetto memorizza le osservazioni di uno o più campioni, insieme a metadati aggiuntivi che descrivono sia le osservazioni (features) sia i campioni (fenotipi).

Un aspetto fondamentale della classe SummarizedExperiment è il coordinamento dei metadati e dei saggi durante la suddivisione. Ad esempio, se si desidera escludere un determinato campione, è possibile farlo sia per i metadati che per i saggi in un'unica operazione, il che garantisce che i metadati e i dati osservati rimangano sincronizzati.

Anatomia di SummarizedExperiment

Il pacchetto SummarizedExperiment contiene due classi:

* **SummarizedExperiment** è un contenitore simile a una matrice in cui le *righe* rappresentano le ***caratteristiche di interesse*** (ad esempio, geni, trascritti, esoni, ecc.) e le *colonne* rappresentano i ***campioni***. Gli oggetti contengono uno o più saggi, ciascuno rappresentato da un oggetto di tipo matrice che a sua volta può essere di tipo numerico o di altro tipo.
  + - * Le righe rappresentano le caratteristiche di interesse. Le informazioni su queste caratteristiche sono memorizzate in un oggetto DataFrame, accessibile con la funzione rowData(). Ogni riga del DataFrame fornisce informazioni sulla caratteristica della riga corrispondente dell'oggetto SummarizedExperiment.
      * Le colonne del DataFrame rappresentano diversi attributi delle caratteristiche di interesse, ad esempio gli ID dei geni o dei trascritti, ecc.
* **RangedSummarizedExperiment** è figlio della classe SummarizedExperiment, il che significa che tutti i metodi di SummarizedExperiment funzionano anche su RangedSummarizedExperiment.

La differenza fondamentale tra le due classi è che le righe di un oggetto RangedSummarizedExperiment rappresentano intervalli genomici di interesse invece di un DataFrame di caratteristiche. Gli intervalli di RangedSummarizedExperiment sono descritti da un oggetto GRanges o un oggetto GRangesList, accessibile con la funzione rowRanges().

> se

class: RangedSummarizedExperiment

dim:

metadata(): ''

assays():

rownames():

rowData names():

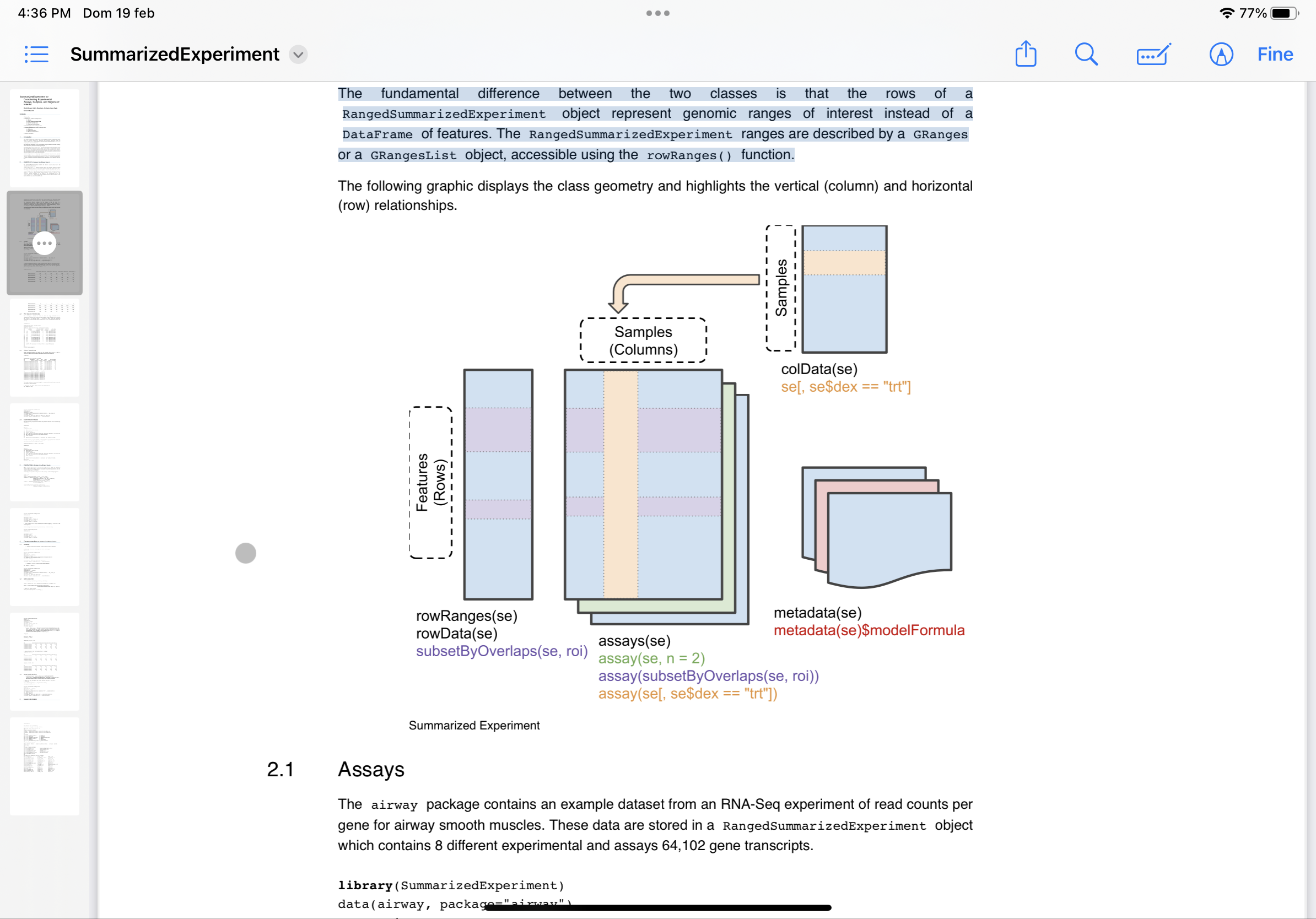
colnames():

colData names():

**Metadata**: metadati che descrivono i metodi sperimentali e i riferimenti di pubblicazione. Possono essere consultati usando metadata()

**Assays**: slot che contiene tutti i dati relativi agli esperimenti eseguiti. Per recuperare i dati degli esperimenti da un oggetto SummarizedExperiment si può usare l'accessor assays(). Un oggetto può avere più dataset di saggi, ciascuno dei quali può essere consultato con l'operatore $.

**RowData** (o RowRanges): dati che rappresentano le caratteristiche di interesse come geni o trascritti. Accessibili con rowData().

**ColData**: metadati che descrivono i campioni, sono accessibili con colData() e rappresentano un DataFrame che può memorizzare un numero qualsiasi di colonne descrittive per ogni riga di campione.

SingleCellExperiment - sce

Oggetto SingleCellExperiment. La classe SingleCellExperiment è un contenitore leggero di Bioconductor per la memorizzazione e la manipolazione dei dati genomici di una singola cellula. Estende la classe SummarizedExperiment e segue convenzioni simili: le righe rappresentano le caratteristiche (geni, trascritti, regioni genomiche) e le colonne rappresentano le cellule (campioni).

La classe SingleCellExperiment è progettata per rappresentare i dati di sequenziamento di una singola cellula. Discende dalla classe RangedSummarizedExperiment e viene utilizzata nello stesso modo. Inoltre, la classe supporta la memorizzazione dei risultati della **riduzione della dimensionalità** (tramite reducedDims) e la memorizzazione di tipi di **caratteristiche alternative** (tramite altExps).

In questa classe, le righe rappresentano caratteristiche genomiche (ad esempio, geni) mentre le colonne rappresentano campioni generati da singole cellule. Diverse quantificazioni (ad esempio, conteggi, CPM, log-espressione) possono essere memorizzate contemporaneamente nello slot dei saggi e i metadati delle righe e delle colonne possono essere allegati utilizzando rispettivamente rowData e colData.

Gli argomenti extra del costrutto (ad esempio, reducedDims altExps) rappresentano le principali estensioni implementate nella classe SingleCellExperiment. Ciò consente di rappresentare in modo coerente e formalizzato le strutture di dati che si incontrano comunemente durante l'analisi dei dati di una singola cellula.

> sce

class: SingleCellExperiment

dim:

metadata(): ''

assays():

rownames:

rowData names():

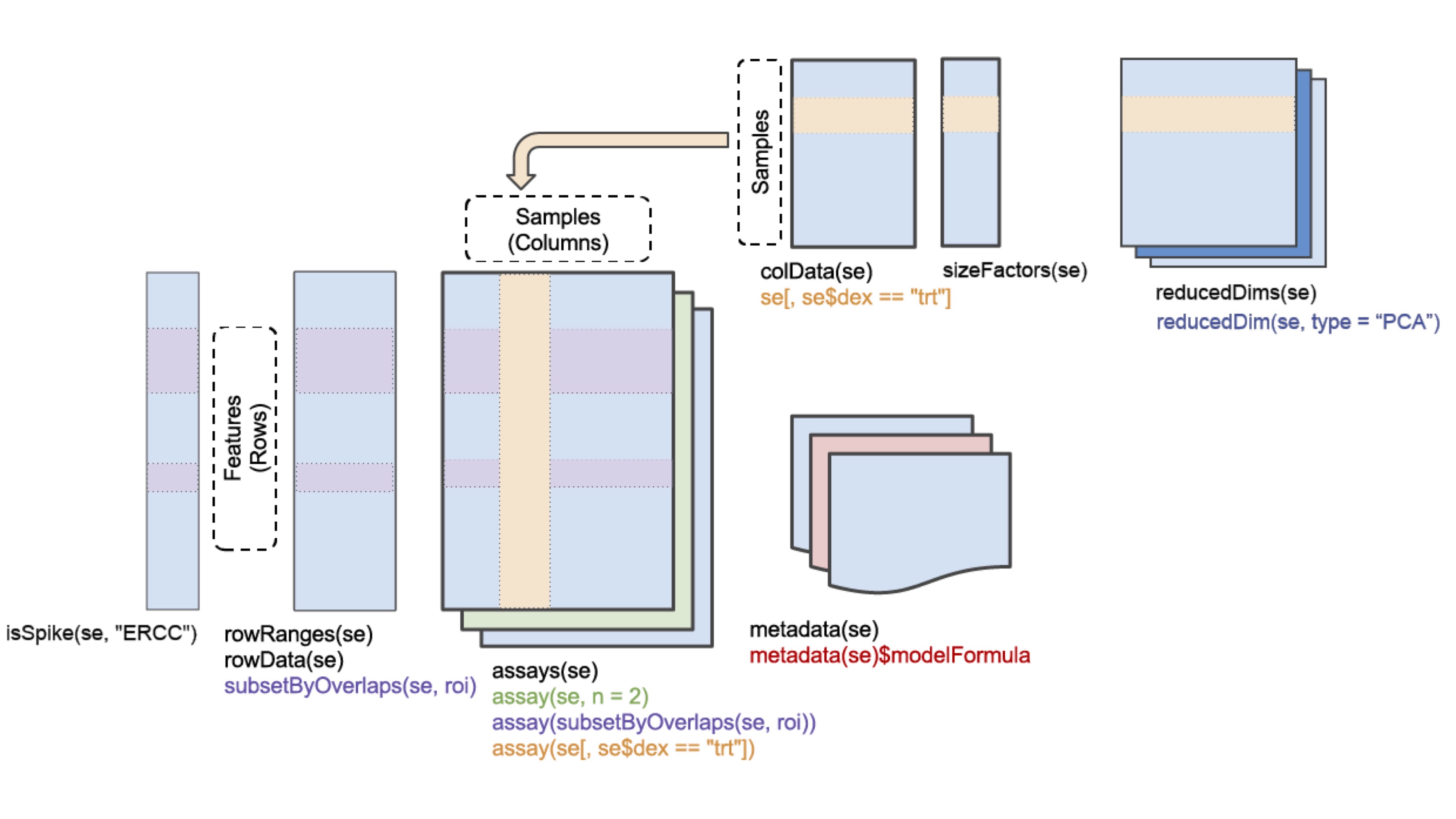
colnames:

colData names():

reducedDimNames():

mainExpName:

altExpNames():



**ReducedDim**: Metodi per ottenere o impostare i risultati della riduzione della dimensionalità in un oggetto SingleCellExperiment. Questi sono tipicamente usati per memorizzare e recuperare rappresentazioni a bassa dimensione di insiemi di dati di cellule singole. Ogni riga di un risultato di riduzione della dimensione corrisponde a una colonna dell'oggetto SingleCellExperiment.

**reducedDim**(x, type, withDimnames=TRUE): Recupera una matrice (o un oggetto simile a una matrice) contenente le coordinate delle dimensioni ridotte per le celle (righe) e le dimensioni (colonne). type è una stringa che specifica il nome del risultato della riduzione di dimensionalità in x da recuperare, oppure uno scalare numerico che specifica l'indice del risultato desiderato, con default alla prima voce se mancante.

Se withDimnames=TRUE, i nomi delle righe della matrice di output vengono sostituiti con i nomi delle colonne di x.

**reducedDimNames**(x): Restituisce un vettore di caratteri contenente i nomi di tutti i risultati della riduzione della dimensionalità di x. È garantito che la lunghezza è uguale al numero di risultati, anche se i nomi possono non essere univoci.

**reducedDims**(x, withDimnames=TRUE): Restituisce una lista di matrici con nome contenente uno o più risultati di riduzione della dimensionalità. Ogni risultato è una matrice (o un oggetto simile a una matrice) con lo stesso numero di righe di ncol(x).

Se withDimnames=TRUE, i nomi delle righe di ogni matrice vengono sostituiti con i nomi delle colonne di x.

**Alternative Experiment**: In alcuni esperimenti, le diverse caratteristiche devono essere normalizzate in modo diverso o avere metadati diversi a livello di riga. Esempi tipici sono i trascritti spike-in negli esperimenti su piastra e i tag anticorpo o CRISPR negli esperimenti CITE-seq. Questi dati non possono essere memorizzati nei saggi principali di SingleCellExperiment. Tuttavia, è comunque possibile memorizzare queste caratteristiche da qualche parte in SingleCellExperiment. Ciò garantisce che i campioni rimangano sincronizzati.

La classe SingleCellExperiment consente di creare "esperimenti alternativi". Gli oggetti di classe SummarizedExperiment annidati sono memorizzati all'interno dell'oggetto SingleCellExperiment x, in modo da garantire che gli oggetti annidati abbiano le stesse colonne nello stesso ordine di quelle presenti in x. Ogni esperimento alternativo deve contenere dati sperimentali e metadati di riga per un insieme distinto di caratteristiche.

SpatialExperiment - spe

La classe SpatialExperiment è una classe S4 di R/Bioconductor per la memorizzazione dei dati di esperimenti di trascrittomica risolti spazialmente (ST). Estende la classe SingleCellExperiment per i dati di una singola cellula per supportare la memorizzazione e il recupero di informazioni aggiuntive da piattaforme ST basate su spot e molecole, comprese le coordinate spaziali, le immagini e i metadati delle immagini.

Un oggetto è composto da:

1. Assays contenenti conteggi di espressione,
2. rowData contenenti informazioni sulle caratteristiche, cioè sui geni,
3. colData contenenti informazioni sugli spot o sulle cellule, compresi i metadati non spaziali e spaziali,
4. spatialCoords contenenti coordinate spaziali
5. imgData contenenti dati di immagine.

Per i dati ST basati su spot (ad esempio, 10x Genomics Visium), viene utilizzato un singolo saggio denominato counts. Per i dati ST basati sulle molecole (ad esempio, seqFISH), è possibile utilizzare due saggi denominati counts e molecules.

> spe

class: SpatialExperiment

dim:

metadata(): ''

assays():

rownames():

rowData names():

colnames():

colData names():

reducedDimNames():

mainExpName:

altExpNames():

spatialCoords names() :

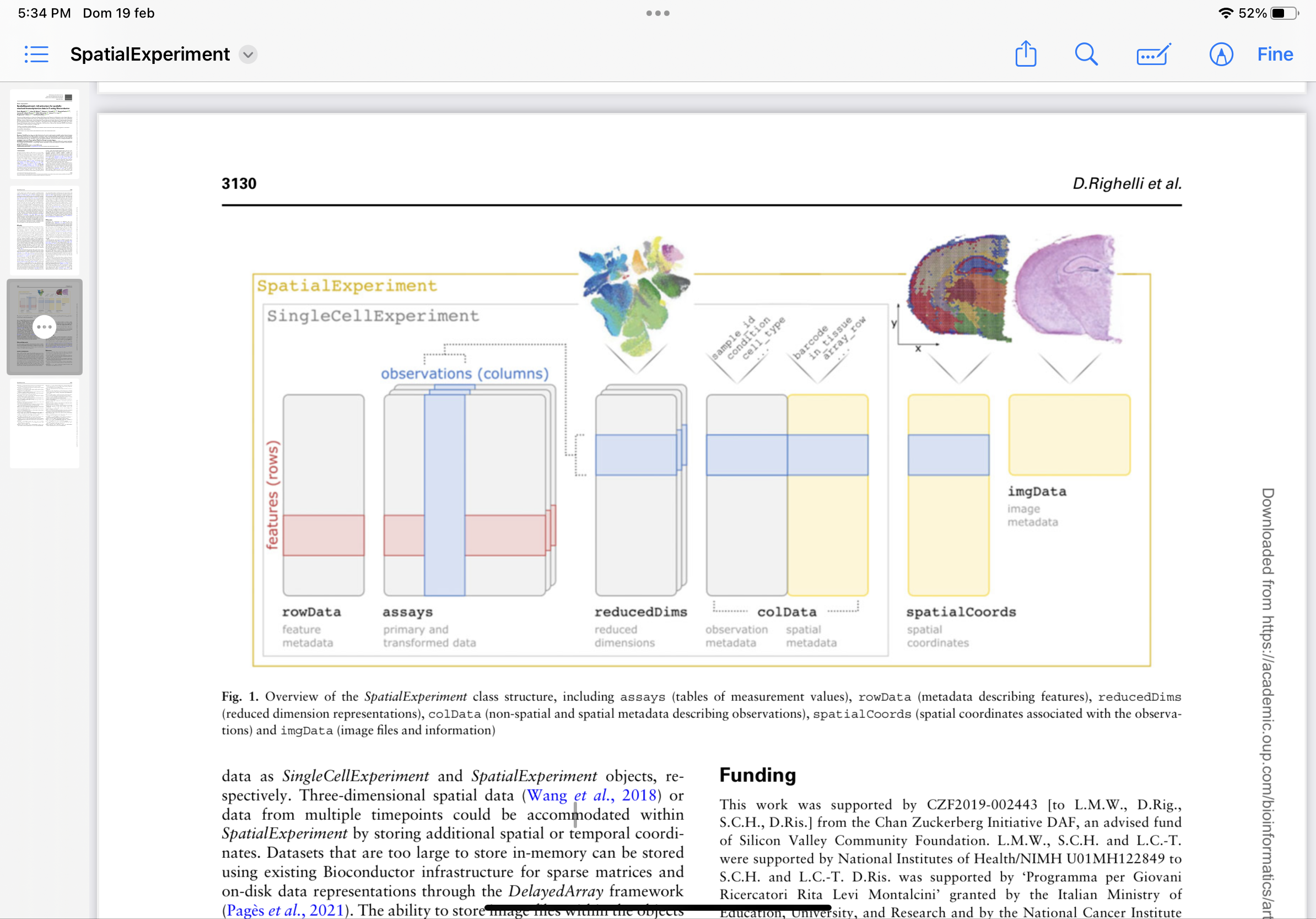
imgData names():

Oltre ai metadati dell'osservazione memorizzati nello slot colData, la classe SpatialExperiment memorizza le coordinate spaziali come:

* **spatialCoords**: una matrice numerica di coordinate spaziali (ad esempio, x e y). Gli spatialCoord sono memorizzati all'interno dello slot int\_colData.

Tutti i dati relativi alle immagini sono memorizzati nel campo imgData di int\_metadata come DataFrame con la seguente struttura:

* ogni riga corrisponde a un'immagine per un dato campione e con un determinato identificativo univoco dell'immagine (ad esempio, le sue risoluzioni)
* per ogni immagine, le colonne specificano
  + - quale ID-campione appartiene all'immagine
    - un ID\_immagine univoco per poter inserire più immagini per un dato campione (ad esempio, con risoluzioni diverse)
    - i dati dell'immagine (un oggetto SpatialImage)
    - scaleFactor che adatta le posizioni dei pixel dell'immagine originale, e dell'immagine a piena risoluzione alle posizioni dei pixel dell'immagine.

Matrici .h5

Visium fornisce matrici nel formato dati gerarchico (HDF5 o H5). H5 è un formato binario che può comprimere e accedere ai dati in modo molto più efficiente rispetto ai formati di testo, il che è particolarmente utile quando si tratta di grandi insiemi di dati.

Il livello superiore del file contiene un singolo gruppo HDF5, chiamato “matrix”, e metadati memorizzati come attributi HDF5. All'interno del gruppo matrice sono presenti set di dati contenenti le dimensioni della matrice, le voci della matrice, nonché le caratteristiche e i codici di cella associati rispettivamente alle righe e alle colonne della matrice.

Le voci della matrice sono memorizzate nel formato Compressed Sparse Column (CSC). Il formato CSC rappresenta la matrice in ordine di colonna maggiore, in modo che ogni codice a barre sia rappresentato da un gruppo contiguo di valori di dati.

Il riferimento alle features è memorizzato come un gruppo HDF5 chiamato features, all'interno del gruppo matrice.

Gerarchia della matrice H5

(root)

└── matrix [HDF5 group]

├── barcodes

├── data

├── indices

├── indptr

├── shape

└── features [HDF5 group]

├─ \_all\_tag\_keys

├─ feature\_type

├─ genome

├─ id

├─ interval [cellranger-arc only]

└─ name

| **Column** | **Type** | **Description** |
| --- | --- | --- |
| barcodes | string | Sequenze di codici a barre e relativi GEM wells (e.g. AAACGGGCAGCTCGAC-1) |
| data | uint32 | Conteggi UMI non nulli in ordine di colonna maggiore |
| indices | uint32 | Indice di riga a base zero dell'elemento corrispondente nei data |
| indptr | uint32 | Indice a base zero nei data / indices dell'inizio di ciascuna colonna, cioè i dati corrispondenti a ciascuna sequenza di codici a barre |
| shape | uint64 | Tupla di (# righe, # colonne) che indica le dimensioni della matrice |