# scShardSplitRef: gene-aware sharding of large chromosomes

Email: irina.kuznetsova@curtin.edu.au

Centre for Crop and Disease Management

GRDC GRAINS RESEARCH & DEVELOPMENT CORPORATION

Curtin University

School of Molecular and Life Sciences, Curtin University

Irina Kuznetsova[1], Luke Pembleton[2], Pasquale Luca Curci[3], Adam Sparks[1]

[1]*Centre for Crop and Disease Management, School of Molecular and Life Sciences, Curtin University, Bentley, WA 6102, Australia*
[2]*Barenbrug, Research and Development, Melbourne, Australia*
[3]*Institute of Biosciences and Bioresources, National Research Council (CNR), Via Amendola 165/A, 70126 Bari, Italy*

## PROBLEM + MOTIVATION

- Single cell multiome ATAC + Gene Expression is used to profile the epigenome and transcriptome simultaneously from the same cells, enabling directly link regulatory DNA accessibility with transcriptional activity at the cellular level
- Cell Ranger ARC software (GTF, FASTA)
- Error for species with large chromosomes/contigs:

  Due to limitations of the BAM index format a contig in the reference FASTA file cannot exceed 536.8 Mb (2^29 bases). If a contig exceeds that size, you will have to split it into smaller contigs and make corresponding modifications to the GTF file

## OBJECTIVE

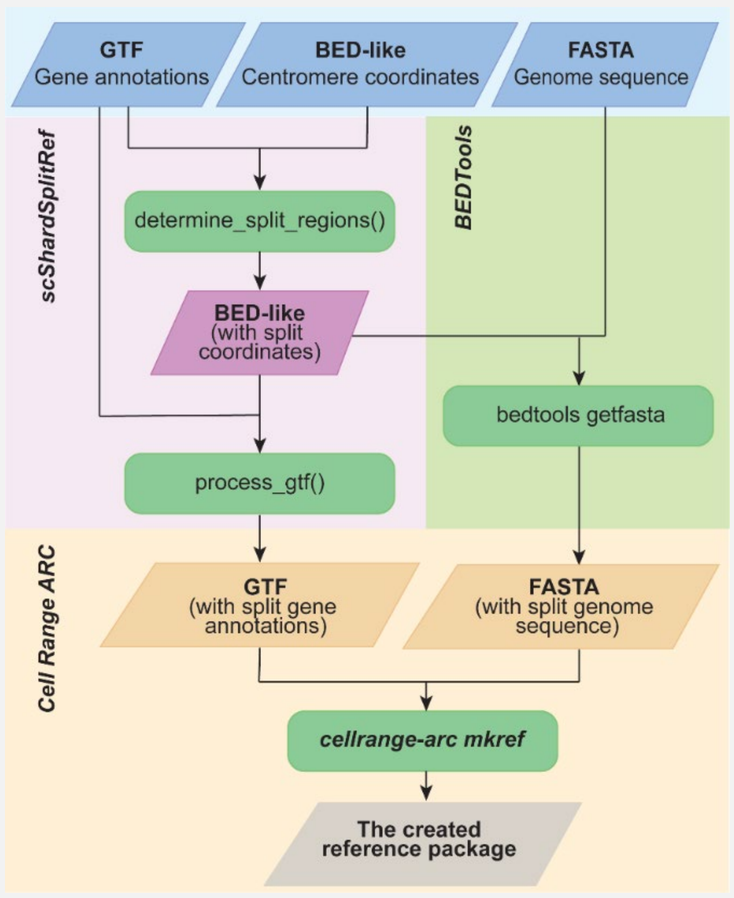
- Develop a solution that provides a reproducible computational approach to fragment large chromosomes into smaller pieces while being aware of gene locations

## THE R PACKAGE / SOLUTION




- **Package:** *scShardSplitRef*
- **Input:** GTF, BED, FASTA input files databases for e.g. Ensembl
- **Output:** GTF, FASTA files containing split chromosomes/contigs
- **What it does:** Fragments large chromosomes/contigs into smaller pieces
- **Core approach:** The core algorithm ensures that chromosome regions remain within the specified Cell Ranger limits and checks that each proposed split coordinate is clear of gene regions according to the specified clearance requirements

## WORKFLOW



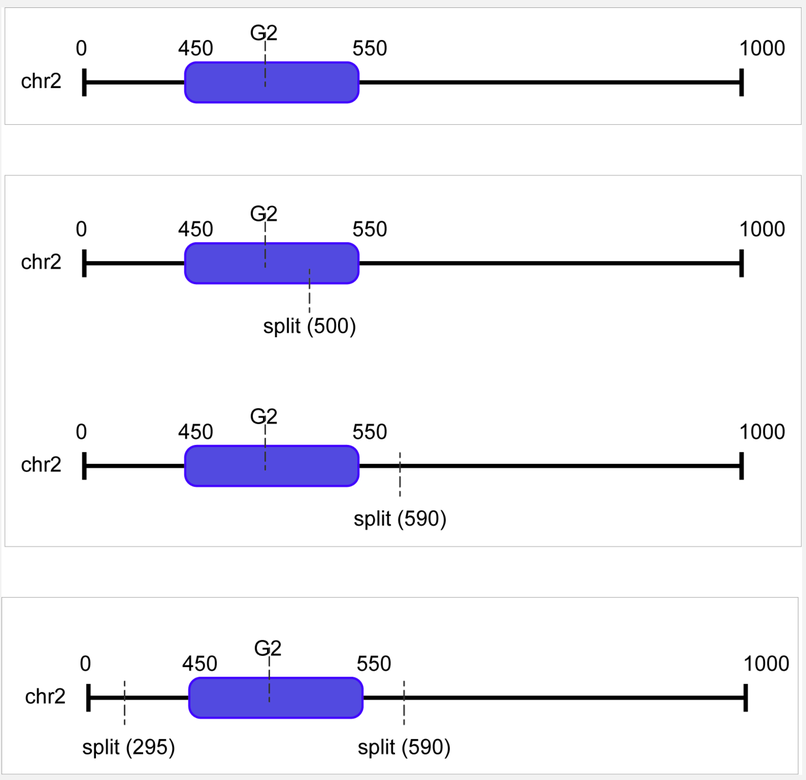

## TOY EXAMPLE

GTF
gene[G2]
450-550

BED
Chr2 0 1000

```
determine_split_regions(bed = path_in_GTF,
gtf = path_in_BED, output_bed = out_path,
                  limit = 540,
                  shift_by = 30,
                  clearance = 40 )
```

- No centromere coordinate is defined in the BED file
- Create provisional even shards based on the `limit`, `ceiling(1000/540)=2`, a split coordinate `500 bp`, which is inside the G2 gene region
- Apply the `shift` parameter, shift rightwards by `30 bp`, a new split coordinate is `530 bp`, still inside the G2
- Shift rightwards by `30 bp`, a new split coordinate is `560 bp`, is outside the G2, but does not pass the `clearance`
- Shift rightwards by `30 bp`, a new split coordinate is `590 bp`, which is outside the G2 region and passes the `clearance`, exceeds the set `limit` of `540 bp`. `ceiling(590/540)=2`, a split coordinate `295 bp`
- Final split coordinates are `295` and `590 bp`




## RESULTS

- CRAN (scShardSplitRef)
- GitHub (scShardSplitRef)
- GPL-3.0 license agreement

- Prepares GTF and FASTA files in compatible format that can be used with Cell Ranger ARC without getting an error

- Tested
  - Barley (*Hordeum vulgare*)
  - Wheat (*Triticum aestivum*)
  - Faba bean (*Vicia faba*)

- Standardised
- Reproducible
- Elegant solution

This work was supported by the Grain Research and Development Corporation (GRDC) through the Analytics for the Australian Grains Industry (AAGI) investment, CUR2210-005OPX (AAGI-CU). This work was supported by resources provided by the Pawsey Supercomputing Research Centre (https://ror.org/04f2f0537) with funding from the Australian Government and the Government of Western Australia. This research used the Australian Research Data Commons (ARDC) ARDC Nectar Research Cloud Service (ARDC Nectar Research Cloud | Australian Research Data Commons). The ARDC is enabled by the Australian Government's National Collaborative Research Infrastructure Strategy (NCRIS).We acknowledge the support provided by the Italian Ministry of University and Research (MUR) under the PRIN 2022 project "Adaptation and Evolution of plants in the urban jungle" (URbEVOLUTION, project code: 2022RYTHE3) for supporting Pasquale Luca Curci's contribution to this work. We acknowledge Dr Fatima Naim (CCDM, Curtin University) and project CUR0023

AAGI Analytics for the Australian Grains Industry

BARENBRUG

pawsey

10X GENOMICS