featureDB – storing and querying genomic annotation

Work in progress …

Arne Müller Preclinical Safety Informatics, Novartis,

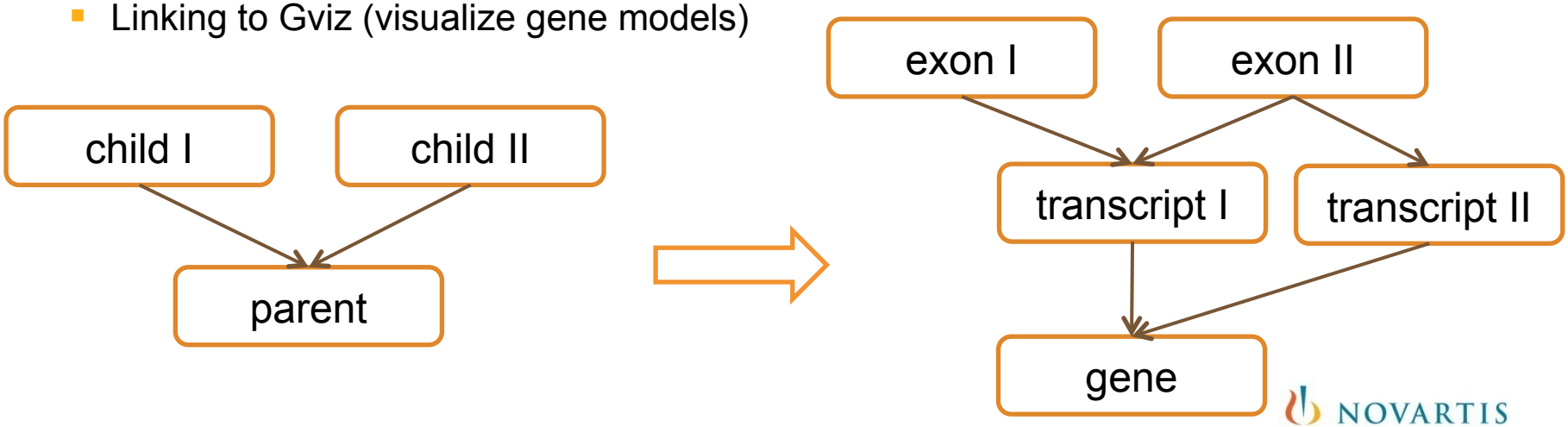
December 14th 2012

Acknowledgements: Florian Hahne + Bioconductor Community

NOVARTIS

featureDB overview and goals

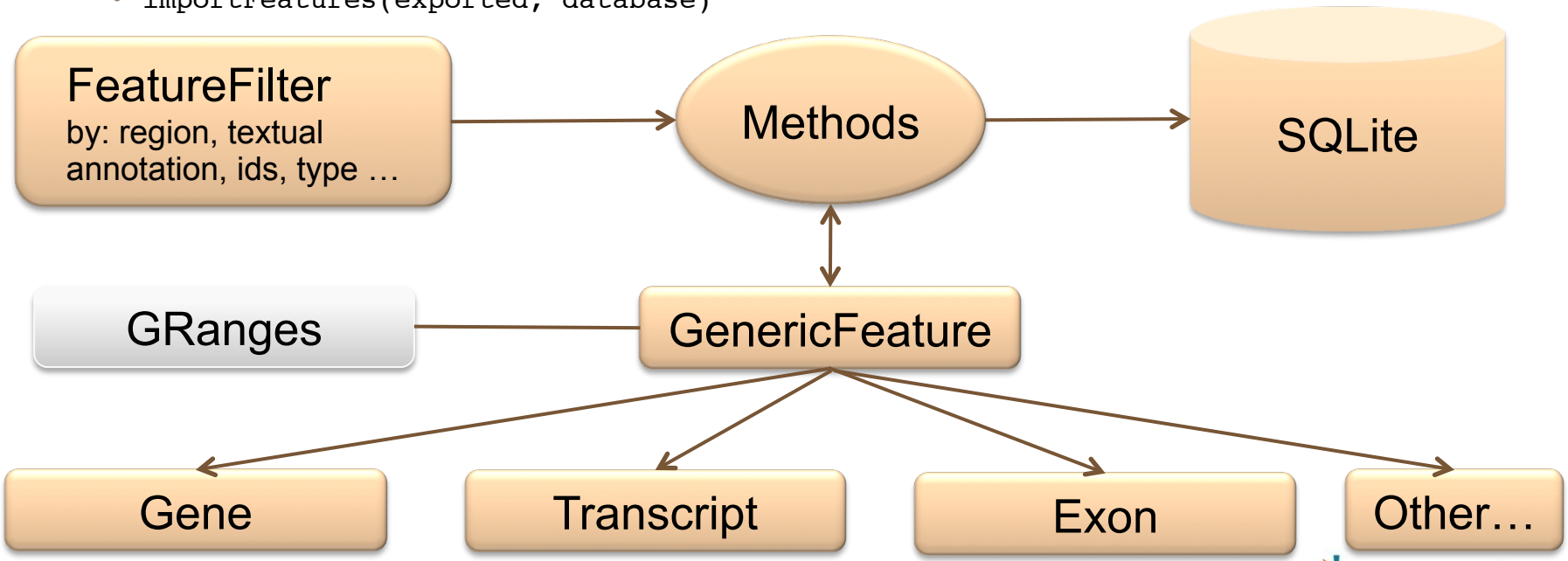
- Re-usable genomic annotations – database back-end for `GRanges` objects (also used by non-R applications)
- Storing, querying and analyzing large sets of annotations
- Hierarchical annotations
- Free text queries: Feature annotation with text: name, synonyms, … searchable *via* a full text index (give me all '*anti-apoptosis*')
- Range queries (GRanges ➔ optimized range queries on the back-end)
- Gene models (genes, transcripts, exons, CDS, UTRs)
- Linking to Gviz (visualize gene models)



NOVARTIS

featureDB architecture (S4 classes/methods)

- Core Methods
  - storeFeatures(object, database)
  - fetchFeature(filter, database)
  - removeFeatures(object, database)
  - fetchParents(object, database)
  - fetchChildren(object, database)
  - linkFeatures(object, database, …)
  - exportFeatures(object, database)
  - importFeatures(exported, database)
- Selected higher level methods
  - fetchUTR(transcript, database)
  - fetchCDS(transcript, database)
  - fetchGeneModel(gene|transcript, database)
  - importGeneModelsFromUCSC(…)
  - importGeneModelsFromBiomart(…)



NOVARTIS

Examples – creating and storing `GenericFeature` objects

```
> library(featureDB)
> library(rtracklayer)
>
> cgi <- as(import.bed("~/tmp/hg19-CGIs.bed"), "GRanges")
> v = values(cgi)
> v$id = v$name
> values(cgi) = v
```

Get some data: `GRanges` with CpG islands
'mandatory `elementMetaData` colums: id' and 'name'

```
> cgiF = createFeature(classname="GenericFeature", genome="hg19", ranges = cgi, name = "CpG
                       Islands", src = "UCSC")
```

Create an `GenericFeature` object

```
> cgiF
Object of class GenericFeature

    PK (Id): 0 (NOT in database!)
    GENOME: hg19
    NAME: CpG Islands
    DESCRIPTION:
    VERSION:
    SOURCE: UCSC
    TIMESTAMP: 2012-11-15 16:53:37
    NUMBER OF RANGES: 2462

GRanges with 2462 ranges and 1 metadata column:
         seqnames                 ranges strand   |        name
            <Rle>              <IRanges>  <Rle>   | <character>
     [1]     chr1       [ 28736,  29810]      *   |    CpG:_116
     [2]     chr1       [135125, 135563]      *   |     CpG:_30
…
> cgiF <- storeFeatures(cgiF, db = "~/tmp/test.db")
> pk(cgiF)
[1] 1
```

Show the object …

Save/store the object

Examples, continued – full text search

```
> genes <- fetchFeatures(FeatureFilter(genome="hg19", src = "UCSC/RefSeq", type = "Gene",
                                       name="ribonucleotide"))
> genes
Object of class Gene

    PK (Id): 1 (in database)
    GENOME: hg19
    NAME: UCSC RefSeq gene model
    DESCRIPTION: known exons, transcripts and genes from RefSeq
    VERSION:
    SOURCE: UCSC/RefSeq
    TIMESTAMP: 2012-11-15 10:21:06
    NUMBER OF RANGES: 4

GRanges with 4 ranges and 7 metadata columns:
        seqnames                 ranges strand |        id      name
           <Rle>              <IRanges>  <Rle> | <factor> <factor>
   2782    chr11 [  4115924,   4160106]      + |     6240     RRM1
   2842     chr2 [ 10262695,  10271546]      + |     6241     RRM2
   7079     chr8 [103216729, 103251346]      - |    50484    RRM2B
  11199     chr2 [216176679, 216214496]      + |      471     ATIC
                                                                      description
                                                                         <factor>
   2782                                                    ribonucleotide reductase M1
   2842                                                    ribonucleotide reductase M2
   7079                                 ribonucleotide reductase M2 B (TP53 inducible)
  11199 5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase
         parentPk       .type   .uniqueId      entrez
        <integer> <character> <character> <character>
   2782      <NA>        Gene        2784        6240
   2842      <NA>        Gene        2844        6241
   7079      <NA>        Gene        7082       50484
  11199      <NA>        Gene       11202         471
```

searches fields: name, description and synonyms

NOVARTIS

Examples, continued – advanced stuff …

```
> tx <- fetchChildren(genes[2])
> tx
Object of class Transcript
…
GRanges with 2 ranges and 9 metadata columns:
        seqnames              ranges strand |           id     name
           <Rle>           <IRanges>  <Rle> |     <factor> <factor>
  35360     chr2 [10262695, 10271546]      + | NM_001165931     RRM2
  29190     chr2 [10262863, 10271546]      + |    NM_001034     RRM2
                      description  parentPk       .type   .uniqueId    cdsEnd
                         <factor> <integer> <character> <character> <integer>
  35360 ribonucleotide reductase M2      2842  Transcript  1458948517  10269513
  29190 ribonucleotide reductase M2      2842  Transcript  1427442084  10269513
        cdsStart        gene
       <integer> <character>
  35360  10262746        6241
  29190  10262926        6241
  …
```

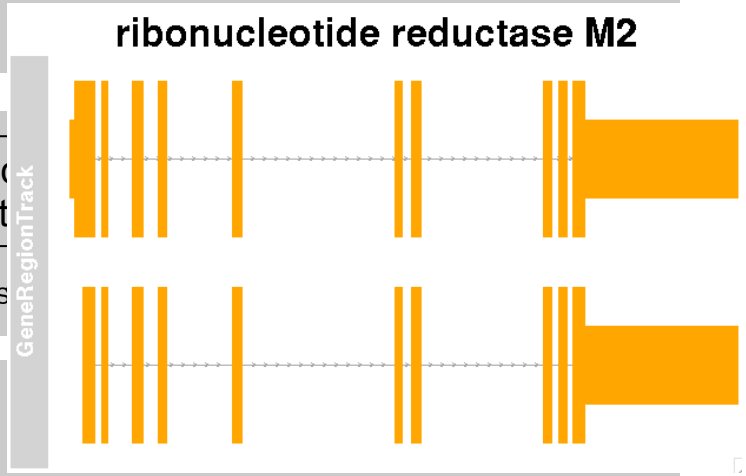
Fetch feature's children (gene ➔transcripts)

```
> cds <- fetchCDS(tx[1], exonic=T)
> library(BSgenome.Hsapiens.UCSC.hg19)
> translate(unlist(getSeq(BSgenome.Hsapiens.UCSC.hg19,
            unlist(cds))))
  450-letter "AAString" instance
seq: MGRVGGMAQPMGRAGAPKPMGRAGSARRGRFKGCWS...LEGKTNFFEKRVGEYQRMGVMS
```

Fetch c
extract

```
> library(Gviz)
> gm <- fetchGeneModel(genes[2])
> plotTracks(gm, main="ribonucleotide reductase M2")
```

Plot gene model



NOVARTIS

Outlook

- Database backend: Add support for PostgreSQL or MySQL
- Full text index: Allow user defined columns
- Performance enhancements
- Convenience: More/better high level methods, updates of objects – any ideas?
  - Note, you have to provide the annotations (`GRanges`) which can he hard work (two methods to fetch gene models from ensembl and ucsc are provided)
- Package under development: Should some of `featureDB` go into the new (or existing) annotation frameworks or become a package?