**Note:** This documents aim to be used as a guide to support the development of lrgext.py, to those intending to modify the script in a future and those looking for a reference of the steps and features to be accomplished during script development in a collaborative environment.

# PROGRAM DEVELOPMENT

### /Need still to include all the things we have done that are included in the script. The final document will contain just the things done and some outlook/

### Initial framework and structure: IGP, Done

### Features:

1. Testing:
   1. LRG file exists and is a readable file: IGP, Done
   2. Check that the xml version format (schema) is right: IGP, Done
   3. Strand test, check direction of strand: IGP, Done
   4. Check if the given gene has an allocated LRG ID: VF, Done
   5. Build test:
      1. Build\_number: check the number of build provided. If more than one, check that coordinates are diff. VF, Done
      2. Build\_coor: check that the start and end of the coordinate for each build are different. VF, Done
   6. Data\_consistency: check that coordinates extracted are numbers: VF, IP
   7. Exon\_coor: check that exons do not overlap: IGP, IP
   8. Exon\_number: check that we get the right number of exons (assert) : IGP, IP
2. Usage:
   1. Extraction of input and outputs from different folders: IGP, Done
   2. Input gene name from command line: VF, Done
   3. Warning if gene not found: VF, Done
3. Parsing: VF + IGP I-P – Check that all our parsing functions are listed
   1. Extracting exon, transcript and protein coordinates per exon: IGP, Done
   2. Extract LRG background info: IGP, Done
   3. Extract gene name: VF, Done
   4. Extract builds info: VF, Done
   5. Extract sequence mismatches between GRCh builds: VF, Done
4. File production:
   1. Create .csv list of gene names and associated LRG IDs: VF, Done
   2. Create tab separated file: IGP, Done
   3. Create .csv file: IGP, Done
   4. Create bed file: IGP, Done
   5. Create .txt build mapper file: VF, Done
   6. Create .csv build differences file: VF, Done

### Disclaimer: IGP, Done

### Dealing with issues and bugs

1. Issue with LRG without protein coordinates in the first exon (e.g. LRG\_62 and LRG\_292): IGP, Done
2. Issue with some genes from list not being called: VF, IP
3. Issue working program not running in all environments: IGP, IP

* **Commenting**:

1. Initial commenting: IGP, Done
2. Update and revision: VF, Done

# Documentation:

### COPYRIGHT:

* 1. Policy checking (University and NHS): IGP, Done
  2. Structure and first drafting: IGP, Done
  3. Completion: IGP, Done
  4. Read proofing: VF, Done

### README (IGP + VF)

1. Document structure: IGP, Done

2. Initial draft: IGP, Done

3. Content:

1. Global Background: VF, IP

2. Usage, installation: IGP, IP (pending on the addition of features)

4. Features description:

* + - 1. Background to group of features: Testing, Parsing & File Generation /VF/
      2. Description of features /IGP + VF as required/

1. create\_repository\_file() : VF
2. initial tests(): IGP + VF (as added by author)
3. handle\_xml(data): IGP
4. get\_gen\_data(data) : VF
5. get\_background(root) : VF
6. get\_build\_info(up\_anno): VF
7. get\_exon\_data(data,gstart,gend,chro,str\_dir) : IGP
8. diff\_data(data): VF
9. coord2file(list\_all\_coord,list4bed): IGP
10. diff2file(build\_data,diff\_data): VF
11. final\_tests(): IGP + VF
12. disclaimer(): IGP
13. Versioning: IGP, Done
14. Disclaimer: IGP, done
15. Final editing and read-proofing: VF, pending completion of documentation

# OUTLOOK

1. Adding new test using 'assert': VF – NOT DONE, UNABLE TO ACHIEVE

2. Creating the name of the file automatically: IGP & VF, Done

3. Including all in the main function (important when program is going to be used as a package): IGP

4. Access file from internet: IGP, deemed as not necessary for now, since LRG files will be downloaded/updated at periodic times to avoid delays due to internet connection