Appendix L.4 Dealing with Psomagen Fastq Files

Often times Psomagen will run a sample across multiple lanes or on multiple runs (ie different flowcells), however when they send us the files they concatenate the FASTQ Files together. This leads to issues downstream in the FASTQ Quality Control and Variant Calling Pipelines.

We must split the FASTQ sequences according to the runs and lanes in order to ensure that the resulting alignments are assigned the correct **Read Group**

Usage:

```
python /share/carvajal-archive/PACKAGES/common-repositories/python/scripts/dna_seq/ValidateFastqReadGroups.py --fastq <gzipped FASTQ fil
```

Example:

```
python /share/carvajal-archive/PACKAGES/common-repositories/python/scripts/dna_seq/ValidateFastqReadGroups.py --fastq GA-255/GA-255_R1.f
```

This will create a file in the RG_Checked Folder with the Flowcell Barcode and Lane information added to the filename. If there aren't multiple read groups in the file the new file will be a symbolic link to the original. Otherwise, it'll create 2 or more files for the read.

Example Output:

1. No extra read groups

`RG_Checked/GA-255_R1_HNCTHDSX2_3.fq.gz` - would be a symbolic link

2. Multiple Read Groups

`RG_Checked/GA-255_R1_HNCTHDSX2_3.fq.gz` - New file FASTQ from Flowcell HNCTHDSX2 Lane 3
`RG_Checked/GA-255_R1_HX534DSG2_1.fq.gz` - New file FASTQ from Flowcell HX534DSG2 Lane 1

Note: I have also created a Jupyter Notebook that implements this splitting for multiple files. Please see -
`/share/carvajal-archive/PACKAGES/common-repositories/python/notebooks/Psomagen_FASTQ_Validation.ipynb`