**Genomics of Natural Populations: Gene Conversion - Analysis Workflow**

File: GeneConversionInversion.Rmd

1. **NG Convert SNP Table to FASTA from Reference**
   1. Purpose: The program takes the SNP table “Chr3\_FB\_SnpTable.tsv” and the reference sequence for Chr3 and generates individual sequences for each strain.
   2. Source: Fortran, NG Convert SNP Table to Fasta from Reference.f
   3. Input File
      1. “*AR\_REF\_Chr3.fas*”, the v3.2 reference sequence for chromosome 3 in *D. pseudoobscura*. “*Chr3\_FB\_SnpTable.tsv*” SNP table from FreeBayes Analysis of 54 strains of *D. pseudoobscura* and 1 strain of the outgroup *D. miranda*.
   4. Output File
      1. “*StrainName\_Chr3.fas*”, where *StrainName* is the name of the strain. THE FASTA files used in the analysis are stored in the data/FASTA Files folder.
2. **NG Convert Fasta to Nexus** 
   1. Purpose: The program converts a set of aligned FASTA sequences and converts them to Nexus format to be able to import the sequences into DNASp for gene conversion analysis.
   2. Source: Fortran, NG Convert Fasta to Nexus.f
   3. Input Files
      1. “*Syntenic\_Regions.txt*”, is a list of 73 regions across the third chromosome of *Drosophila pseudoobscura* with the name of the region, the beginning and end of the cytogenetic region, and the beginning and end coordinate in the version 3.2 assembly.
      2. “*snp\_Dpse\_All.fof*”, is a list of the 55 fasta sequences for chromosome three of 54 *D. pseudoobscura* strains and the outgroup strain *D. miranda*.
      3. “*X\_Chr3.fas*”, is the sequence file where X is the strain name.
   4. Output File
      1. *“Chr3\_Syn\_Reg\_xxx.nex*” is the name of the output nexus file where xxx is the name of the subregion [01a-14h]. These files are stored in the data/Nexus Files folder.
3. **Syntenic Block Rearrangement Analysis**
   1. Purpose: The R code chunk determines the order of the 73 syntenic subregions in the different third chromosome gene arrangements.
   2. Source: R code chunk in GeneConversionInversion
   3. Input Files
      1. “*AR\_Syntenic\_Regions.txt*”, region order in the AR reference genome.
      2. “*ST\_Ord.txt*”, subregion order in ST relative to AR
      3. “PP\_Ord.txt”, subregion order in PP relative to ST
      4. “*HY\_Ord.txt*” subregion order in HY relative to ST
      5. “*SC\_Ord.txt*”, subregion order in SC relative to HY
      6. “*CU\_Ord.txt*”, subregion order in CU relative to SC
      7. “CH\_Ord.txt”, subregion order in CH relative to SC
      8. “*TL\_Ord.txt*”, subregion order in TL relative to SC
   4. Output Files in the output folder
      1. “*AR\_coordinates.csv*”, where *arr* is the gene arrangement name (AR, ST, PP, HY, SC, CU, CH, or TL).
      2. “*ST\_coordinates.csv*”, where *arr* is the gene arrangement name (AR, ST, PP, HY, SC, CU, CH, or TL).
      3. “*PP\_coordinates.csv*”, where *arr* is the gene arrangement name (AR, ST, PP, HY, SC, CU, CH, or TL).
      4. “*HY\_coordinates.csv*”, where *arr* is the gene arrangement name (AR, ST, PP, HY, SC, CU, CH, or TL).
      5. “*SC\_coordinates.csv*”, where *arr* is the gene arrangement name (AR, ST, PP, HY, SC, CU, CH, or TL).
      6. “*CU\_coordinates.csv*”, where *arr* is the gene arrangement name (AR, ST, PP, HY, SC, CU, CH, or TL).
      7. “*CH\_coordinates.csv*”, where *arr* is the gene arrangement name (AR, ST, PP, HY, SC, CU, CH, or TL).
      8. “*TL\_coordinates.csv*”, where *arr* is the gene arrangement name (AR, ST, PP, HY, SC, CU, CH, or TL).
      9. The output files each have three columns [Subregion, Subregion begin coordinate, Subregion end coordinate].
4. **DNASP Analysis Gene Conversion Analysis**
   1. Method: The 73 nexus files of aligned sequences were imported into DNASP. The Gene Conversion Analysis within DNASP was run on pairs gene arrangement subsamples.
   2. Output Files in the data folder
      1. Results of the analysis were stored in text files named “*GeneConversion\_xxx\_arr1\_arr2.txt*” where *xxx* is the subregion name, *arr1* is the name of arrangement 1 and *arr2* is the name of arrangement 2. A total of 1,095 text files were generated from the pairwise analysis of six arrangements, 15 per region, and 73 regions. Each analysis excluded nucleotide sites with alignment gaps and provided lists of conversion tracts with the strain, the beginning and end coordinates, tract length excluding gaps, mean value of y, number of sites with information, and a list of sites with associated value of y.
5. **Concatenate DNASP Gene Conversion Data**
   1. Purpose: Uses the output from DNASP output files from the gene conversion analysis (step 4) and concatenates gene conversion tract information into a single file “*GeneConversion\_Tracts.tsv*”, compiles a statistics file “*GeneConversion\_Stats.tsv*”, and compiles a list of informative sites “*GeneConversion\_Sites.tsv*”
   2. Source: PERL, Extract DNASP Gene Conversion.pl
   3. Input File
      1. “*file\_list.fof*”, is a list of 1,095 file names that were output from DNASP. The list of names reflects 15 pairwise comparisons of six gene arrangements (AR, ST, PP, CH, TL, CU) for 73 regions of 181,398 – 440,325 nucleotides with 49 regions being 250,000 nucleotides.
   4. Output Files
      1. “*GeneConversion\_Tracts.tsv*”, is a tab separated values file that lists all gene conversion tracts from all pairwise comparisons of the six gene arrangements. The file has six columns [**GC\_Event**, **GC\_File**, **GC\_Strain**, **GC\_Beg**, **GC\_End**, **GC\_Len**], where GC\_Event is the number of the event; GC\_File is the filename with the data; GC\_Strain is the strain with the gene conversion event; GC\_Beg is the first nucleotide of the gene conversion tract; GC\_End is the last nucleotide of the gene conversion tract; GC\_Len is the length of the gene conversion tract. The file is stored in the data folder.
      2. “*GeneConversions\_Stats.tsv*”, is a tab separated values file that lists the statistics for the 15 pairwise comparisons of the six gene arrangements for the 73 regions for a total of 1,095 analyses. The file has nine columns [**GC\_File**, filename with the data; **GC\_Pop1**, gene arrangement 1; **GC\_Pop1\_No**, number of gene arrangement 1; **GC\_Pop2**, gene arrangement 2; **GC\_Pop2\_No**, number of gene arrangement 2; **GC\_Tract\_No**, number of gene conversion tracts; **GC\_Sites**, number of informative sites; **GC\_Psi**, mean value of Psi for the informative sites; **GC\_Phi**, mean value of Phi for the observed data] where GC\_File is the filename with the data; GC\_Pop1 is the name of gene arrangement 1; GC\_Pop1\_No is the sample size of gene arrangement 1; GC\_Pop2 is the name of gene arrangement 2; GC\_Pop2\_No, is the sample size of gene arrangement 2; GC\_Tract\_No is the number of gene conversion tracts; GC\_Sites, is the number of informative sites; GC\_Psi is the mean value of Psi for the informative sites; GC\_Phi is the mean value of Phi for the observed data. The file is stored in the data folder.
      3. “*GeneConversion\_Sites\_Regxxx.tsv*”, is a tab separated values file that lists the informative sites used in the gene conversion detection analysis for the xx Region for Regions [1-14]. The file has six columns [**GC\_File**, filename with the data; **GC\_Pop1**, gene arrangement 1; **GC\_Pop2**, gene arrangement 2; **GC\_SiteNo**, nucleotide position of the informative site in the data subset; **GC\_SitePsi**, value of Psi for the informative site; **TotalSites**, total number of nucleotide sites in the subregion] where GC\_File is the filename with the data; GC\_Pop1 is the name of gene arrangement 1; GC\_Pop2 is the name of gene arrangement 2; GC\_SiteNo is the nucleotide position of the informative site in the data subset; GC\_SitePsi is the value of Psi for the informative site; TotalSites is the total number of nucleotide sites in the subregion. The file is stored in the data folder.
      4. “*GeneConversion\_Check.tsv*” - a table separated value file that checks the input file name matches the contents of the file. The file is stored in the data folder.
6. **Merge Psi and Filter Data**
   1. This R code chunk merges the estimate of Psi with the GC\_Tracts Data. A new variable “GC\_Tract\_Count” that counts the number of pairwise comparisons that identify a particular tract in the “Tract” variable. The table is sorted by “Tract”. A new variable “exclude” is added do that only a single instance of each “Tract” is used. Non-redundant tracts are identified with “N” while redundant tracts are identified with :Y” in “exclude”.
   2. Source: R code chunk in InversionGC.Rmd
   3. Input Files in the data folder
      1. “*GeneConversion\_Stats.tsv*” from step 5
      2. “*GeneConversion\_Tracts.tsv*” from step 5.
   4. Output Files
      1. “*Tract\_Count.csv*”, number of pairwise comparisons identifying each gene conversion tract. The file is stored in the data folder.
      2. “*GeneConversion\_Tracts\_Processed.csv*”, updated “*GeneConversion\_Tracts.tsv*” file with “Psi” and “exclude” variables added. The file is stored in the data folder.
      3. “*GC\_List.tsv*”, list of the unique gene conversion tracts with seven columns [**Don\_Arr**, **Rec\_Arr**, **Region**, **GC\_Len**, **Len**, **Psi**, **exclude**], where Don\_Arr is the donor arrangement; Rec\_Arr is the recipient arrangement; GC\_Len is the gene conversion tract length; Psi is the mean probability of SNP sites being informative for gene conversion analysis; and exclude is the whether the tract should be filtered out or not. The file is stored in the data folder.
      4. “*GC\_arr\_Tracts.tsv*”, list of gene conversion tracts for the arr arrangement. The file has eight columns [**Subregion**. **Subregion Begin Coordinate**, **Subregion End Coordinate**, **Region Begin Coordinate**, **Region End Coordinate**, **Region**, **Chromosome Beg Coordinate**, **Chromosome End Coordinate**]. The first line has column labels. The remaining lines contain the conversion tract information. The file is stored in the data folder.
7. **NG Gene Conversion Tract Format File**
   1. Purpose: This program sets up text files with observed gene conversion tract length data for the Fortran program NG Betran GC ML, which estimates actual gene conversion parameters (Betran *et al.* 1997).
   2. Source: Fortran, NG Gene Conversion Tract Format File.f
   3. Input File in the data folder
      1. “*GC\_List.tsv*”, is a tab separated value file with all gene conversion tracts for each inversion pairs. This file was generated in the Merge Psi and Filter Data chunk of GeneConversionInversion.Rmd file. The file has six columns [Donor Arrangement, Recipient Arrangement, Syntenic Block Region, Observed Gene Conversion Tract Length, Psi, exclude] and each row is the data for a particular tract length.
   4. Output Files in the data folder
      1. “*GCT\_arr1\_arr2\_reg.txt*” where *arr1* is the donor arrangement, *arr2* is the recipient arrangement, and *reg* is the syntenic block region. First line of the file is the number of gene conversion tracts for the donor and recipient arrangement and the region. Remaining lines have two columns [Observed Gene Conversion Tract Length, Psi Value]. The file is stored in the data folder.
      2. “*GC.fof*” is a file of filenames that contain gene conversion data. The filenames are (GCT\_arr1\_arr2\_reg.txt” where *arr1* is the donor arrangement, *arr2* is the recipient arrangement, and *reg* is the syntenic block region. The file is stored in the data folder.
8. **NG Betran GC ML**
   1. Purpose: This program estimates maximum likelihood gene conversion parameters from a list of observed gene conversion tract lengths from DNASP output. We used the maximum likelihood approach of (Betran *et al.* 1997) using equations 4, 6, and 8. The second derivative of the likelihood equation was used to estimate the asymptotic variance. The value of phi is iterated from 0.98 to 1.00 in increments of 0.000001.
   2. Source: Fortran, NG Betran GC ML.f
   3. Input Files in the data foler
      1. “*GC.fof*” is a file of filenames that contain gene conversion data. Files are opened sequentially and analyzed to estimate gene conversion parameters. The files were created in step 7/
      2. “*GCT\_arr1\_arr2\_reg.txt*”, where *arr1* is the donor arrangement, *arr2* is the recipient arrangement, and *reg* is the syntenic block region. First line of the file is the number of gene conversion tracts for the donor and recipient arrangement and the region. Remaining lines have two columns [Observed Gene Conversion Tract Length, Psi Value].
   4. Output Files in the data folder
      1. “*GC\_LH\_arr1\_arr2\_reg.csv*”, is a comma separated value file that outputs the three columns [Iteration; Phi, value of the phi that is being used to estimate the likelihood value; lnL, likelihood value estimated based on the value of phi].
      2. “A\_Gene\_Conv\_Estimate.txt”, is a text file with the output of the maximum likelihood analysis. Each line represents the values from each file of gene conversion data. The file has 12 columns [Don, donor arrangement; Rec, recipient arrangement; Reg, syntenic block region 1-14; Phi, maximum likelihood estimate of phi; LowLim(N), lower limit on the gene conversion tract length based on maximum likelihood estimate of phi minus two times the standard deviation; Expect(N), gene conversion tract length based on the maximum likelihood estimate of phi (1/(1-phi)); UpLim(N), upper limit on the gene conversion tract length based on maximum likelihood estimate of phi plus two times the standard deviation; Prob(UE), probability of an undetected gene conversion event equation 9 in Betran *et al*. (1997); Obs#(CT), observed number of conversion tracts (k) from DNASP; Exp#(CT), expected number of conversion tracts = k/P(L>=2) see page 95 column 1 in Betran *et al*. (1997); Prob(TS), probability of a transferred site = (Exp#(CT)\*Expect(N))/(number of sequences\*number of nucleotides) see page 95 column 2 in Betran *et al* (1997)].
9. **NG Build Coordinate Spreadsheet**
   1. Purpose: Reorganizes the “*Coordinate.txt*” file into a new “Coordinates\_List.csv” file.
   2. Source: Fortran, NG Build Coordinate Spreadsheet.f
   3. Input File
      1. “*Coordinates.txt*”, Line 1 - Transcript Name, Orientation; Line 2 – Exon number; Line 3 – First nucleotide of exon, Last nucleotide of exon; subsequent lines for each exon; Format repeats for all transcripts.
   4. Output File
      1. “*Coordinates\_List.csv*”, is a comma separated value file that lists each transcript exon on a separate line. Columns [Transcript\_ID, name of the transcript; Ori, orientation of the exon; Exon, exon number; Pos\_Beg, first nucleotide of the exon; Pos\_End, last nucleotide of the exon].
   5. Note
      1. The “*Coordinates\_List.csv*” was imported into Excel and saved as “2019\_02\_02 Coordinates\_List.xlsx.” This file has 13 tabs [1. BP, 2. Coordinates\_List, 3. Culled Coordinates\_List, 4. AR\_List, 5. AR\_Outlier List, 6. ST List, 7. ST Outlier List, 8. PP List, 9. PP Outlier List, 10. CH List, 11. CH Outlier List, 12. TL List, 13. TL Outlier List]. 1. BP, provides the starting nucleotide for each syntenic block in each arrangement. 2. Coordinates\_List, imported data from the “Coordinates\_List.csv” file. The last column “Exclude Transcript” indicates if the exon excluded because it overlaps with another transcript. 3. Culled Coordinates\_List, list of exons with the redundant exons removed. Translates the coordinates of the exons from the AR reference sequence to that of ST, PP, CH, TL, and CU. AR\_Ord, order of exons in AR; AR\_Ori, orientation of exon in AR; AR\_Pos\_Beg, first nucleotide of the exon in AR; AR\_Pos\_End, last nucleotide of the exon in AR; AR\_Outlier, indicates if the exon is part of an outlier gene in AR. The same information is given for ST, PP, CH, TL, and CU. 4. AR\_List, Lists the non-redundant exons for AR. New Columns [Outliers: 1, non-outlier; 2, outlier; Overlap: No, Yes]. 5. AR Outlier List, only includes the coordinates of outlier transcripts from beginning to end. Each arrangement has a “List” and an “Outlier List” file. The “5. AR Outlier List” was exported to “GC\_AR\_Outlier\_List.txt” and each of the other arrangements had a similar text file with coordinates of the outlier transcripts. The text file has four columns, Column 1, Transcript Name; Column 2, Midpoint nucleotide coordinate; Column 3, Beginning nucleotide coordinate; Column 4, End nucleotide coordinate.
10. **NG Outlier Gene Map**
    1. Purpose: Plots the location of outlier genes on the map of chromosome 3. Chromosome 3 is represented as vertical rectangle and outlier genes are represented as dots on the map. The top of the page is the 5’ end to 3’ end at the bottom of the page.
    2. Source: Fortran, NG Outlier Gene Map.f
    3. Input File
       1. Console: Gene arrangement name *arr* where *arr* is the two-letter inversion code [AR ST PP CH TL CU]
       2. “*GC\_arr\_Outlier\_List.txt*” This file has four columns [**Transcript**, **Midpoint**, **Beg**, **End**].
    4. Output File
       1. “*GC\_arr\_Outlier\_Plot.ps*”
    5. Note
       1. Each postscript file (\*.ps) for each gene arrangement is converted to a pdf file with Adobe Distiller. The pdf files for the arrangements are imported and merged into Adobe Illustrator to show the locations of the outlier genes on the third chromosomal map.
11. **NG Gene Conversion Tract Map**
    1. Purpose: Plots gene conversion events to a chromosomal map in two ways, as tracts across the map and as a histogram of the number of tracts that cover a particular nucleotide. This is an exploratory tool that allows one to choose an arrangement and provide beginning and end coordinates leading to the plot of gene conversion events and genes locations.
    2. Source: Fortran, NG Gene Conversion Tract Map.f
    3. Input Files
       1. Console: Gene arrangement name (*arr*) and beginning (*beg*) and end (*end*) coordinates for the plot.
       2. “*GC\_arr\_Tracts.tsv*” from step 5 lists the beginning and end nucleotide of each gene conversion tract for the *arr*.
       3. “*GC\_arr\_Transcripts\_List.tsv*”, lists the exons for all transcripts in *arr* coming from the “*arr* List” tab from “2019\_02\_02 Coordinates\_List.xlsx.”
    4. Output Files
       1. “*GC\_arr\_Tracts\_Output.csv*”, lists the intervals with gene conversion tracts. Columns [Region, number of the region having at least one gene conversion tract covering the nucleotide; Beg, first nucleotide of the region; End, last nucleotide of the region; Count, mean number of gene conversion tract coverage for the region]
       2. “*GC\_arr\_Transcript\_Tract\_Depth.csv*”, lists the mean gene conversion coverage for each exon of all transcripts. The file has six columns [**Transcript**, **Exon**, **Beg**, **End**, **Outlier**, **Mean\_GC**].Transcript is the transcript name; Exon is the number of the exon; Beg is the first nucleotide of the exon; End is the last nucleotide of the exon; Outlier is the outlier status of the transcript where 1 is non-outlier and 2 is outlier; Mean\_GC is the mean gene conversion coverage per nucleotide.
       3. “*GC\_arr\_Plot\_beg\_end.ps*” is a postscript file that plots the gene conversion tracts to an interval of the chromosome.
    5. Note
       1. The postscript file (\*.ps) is converted to a pdf file with Adobe Distiller. The pdf file is imported into Adobe Illustrator to show the exons of transcripts, non-outliers in black and outliers in blue, gene conversion tract intervals, and a histogram of gene conversion coverage.
12. **NG Reg Don Rec Permute**
    1. Purpose: This program a permutation test to determine if the frequency of donor to recipient gene conversion tracts is independent among the 14 regions and overall. A total of 10,000 random permutations are used.
    2. Source: NG Reg Don Rec Permute.f
    3. Input:
       1. “*GC\_Region\_Donor\_Recipient.tsv*” a tab-separated file with three columns [Region, Donor Arrangement, Recipient Arrangement].
    4. Output File
       1. “*GC\_Region\_Donor\_Recipient\_Summary.txt*” For each region and the overall total, the observed donor to recipient frequencies are presented, the minimum frequency across the permutations, the maximum frequency across the permutations, the probability of permutation values less than the observed, and the probability of permutations greater than the observed. The random seed iseed is printed.
13. **NG GC Tract Length Region Permutation**
    1. Purpose: This program uses a permutation test to determine whether the median value in each of the 14 regions represents a deficiency or an excess compared to randomly permuted gene conversion tract. This program randomly permutes the assignment of tracts to region.
    2. Source: Fortran, NG GC Tract Length Region Permutation.f
    3. Input:
       1. Console: Gene arrangement name *arr*, where *arr* is the two letter inversion code [AR ST PP CH TL CU]
       2. *“GC\_arr\_Tracts.tsv*”, from step 5 a tab separated value file with a list of gene conversion tracts, one tract per row.
    4. Output Files
       1. “*GC\_arr\_ Tract\_Length\_Permut.txt*”, a text file with the results of the random permutation test. A table is presented with four columns [Region, Observed, Probability(Permuted <Obs), Probability(Permuted <Obs)] and fourteen rows, one for each region. The probability of mean GC coverage for outlier genes bases on the number of random permutations less than or equal to the observed values is given. Finally, the minimum and maximum value of 10,000 random permutations is presented for each region. A file is generated for each arrangement [AR, ST, PP, CH, TL]. The random seed iseed is printed.
14. **NG Gene Conversion Permutation Test**
    1. Purpose: This program uses a permutation test to determine whether outlier genes have different levels of gene conversion tract coverage than non-outlier genes. The set of gene conversion tracts is mapped to the third chromosome, then a histogram of coverage for each nucleotide is determined. The mean gene conversion tract coverage is estimated for outlier genes and non-outlier genes. This program randomly permutes the assignment of outliers to transcripts.
    2. Source: Fortran, NG Gene Conversion Permutation Test.f
    3. Input
       1. Console: Gene arrangement name *arr*, where *arr* is the two letter inversion code [AR ST PP CH TL CU]
       2. “*GC\_arr\_Tracts.tsv*”, from step 5 a tab separated value file with a list of gene conversion tracts, one tract per row.
       3. “*GC\_arr\_Transcripts\_List.tsv*”, lists the exons for all transcripts in *arr* coming from the “*arr* List” tab from “2019\_02\_02 Coordinates\_List.xlsx.”
    4. Output Files
       1. “*GC\_arr\_Tracts\_Output\_Permut.csv*”, lists the intervals with gene conversion tracts. The file has four columns [**Region**, **Beg**, **End**, **Count**] where Region is the number of the region having at least one gene conversion tract covering the nucleotide; Beg is the first nucleotide of the region; End is the last nucleotide of the region; Count is the mean number of gene conversion tract coverage for the region.
       2. “*GC\_arr\_Transcript\_Tract\_Depth\_Permut.csv*”, lists the mean gene conversion coverage for each exon of all transcripts. The file has six columns [**Transcript**, **Exon**, **Beg**, **End**, **Outlier**, **Mean\_GC**] where Transcript is the transcript name; Exon is the number of the exon; Beg is the first nucleotide of the exon; End is the last nucleotide of the exon; Outlier is the outlier status of the transcript where 1 is non-outlier and 2 is outlier; Mean\_GC, mean gene conversion coverage per nucleotide.
       3. “*GC\_arr\_Random\_Permutation\_Data.csv*”, a comma separated value file with the mean coverage for the outlier genes in each of 10,000 random permutations.
       4. “*GC\_arr\_Outlier\_Tract\_Depth\_Permute.txt*”, a text file with the results of the random permutation test. A table is presented with three columns [Mean GC Bases, Variance GC Bases, Nucleotides] and two rows {Non-Outlier, Outlier]. The probability of mean GC coverage for outlier genes bases on the number of random permutations less than or equal to the observed values is given. Finally, the minimum and maximum value of 10,000 random permutations is presented. A file is generated for each arrangement [AR, ST, PP, CH, TL]. The random seed iseed is printed.
15. **NG Gene Conversion Shuffle Tracts**
    1. Purpose: This program uses shuffled gene conversion tracts to determine whether outlier genes have different amounts of gene conversion than non-outlier genes.
    2. Source: Fortran, NG Gene Conversion Shuffle Tracts.f
    3. Input files
       1. Console: Gene arrangement name *arr* where *arr* is the two letter inversion code [AR ST PP CH TL CU]
       2. “*arr\_BP.txt*”, This file has the three proximal, inverted, and distal intervals for the derived arrangement. Line one has the number of intervals, line two has the coordinates for the proximal interval, line three has the coordinates for the inverted interval, and line four has the coordinates for the distal interval.
       3. “*GC\_arr\_Tracts.tsv*”, from step 6 a tab separated value file. The first line has column labels and the remaining lines are a list of gene conversion tracts, one tract per row.
       4. “*GC\_arr\_Transcripts\_List.tsv*”, lists the exons for all transcripts in *arr* coming from the “*arr* List” tab from “2019\_02\_02 Coordinates\_List.xlsx.”
    4. Output Files
       1. “*GC\_arr\_Tracts\_Output\_Shuff.csv*”, The file has four columns [**Region**, **Beg**, **End**, **Count**] where Region is the number of the region having at least one gene conversion tract covering the nucleotide; Beg is the first nucleotide of the region; End is the last nucleotide of the region; Count is the mean number of gene conversion tract coverage for the region.
       2. “*GC\_arr\_Transcript\_Tract\_Depth\_Shuff.csv*”, lists the mean gene conversion coverage for each exon of all transcripts. The file has six columns [**Transcript**, **Exon**, **Beg**, **End**, **Outlier**, **Mean\_GC**] where Transcript is the transcript name; Exon is the number of the exon; Beg is the first nucleotide of the exon; End is the last nucleotide of the exon; Outlier is the outlier status of the transcript where 1 is non-outlier and 2 is outlier; Mean\_GC, mean gene conversion coverage per nucleotide.
       3. “*GC\_arr\_Random\_Permutation\_Data\_Shuff.csv*”, a comma separated value file with the mean coverage for the outlier genes in each of 10,000 random permutations.
       4. “*GC\_arr\_Outlier\_Tract\_Depth\_Shuff.txt*”, a text file with the results of the gene conversion tract shuffling test. A table is presented with three columns [Mean GC Bases, Variance GC Bases, Nucleotides] and two rows {Non-Outlier, Outlier]. The probability of mean GC coverage for outlier genes bases on the number of random permutations less than or equal to the observed values is given. Finally, the minimum and maximum value of 10,000 random permutations is presented. A file is generated for each arrangement [AR, ST, PP, CH, TL]. The random seed iseed is printed.
16. **NG Gene Conversion Shuffle Tracts ML**
    1. Purpose: Uses shuffled tracts to determine whether outlier genes have different amounts of gene conversion then non-outlier genes. The tract length distribution is based on the ML estimates from the Betran et al. (1997) paper
    2. Source: Fortran, NG Gene Conversion Shuffle Tracts ML.f
    3. Input Files
       1. Console: Gene arrangement name *arr* where *arr* is the two letter inversion code [AR ST PP CH TL CU]
       2. “*arr\_MinMax.tsv*”, This file has the beginning and end coordinates for the 14 regions in arr arrangement.
       3. “*GC\_arr\_Tracts.tsv*”, from step 6 a tab separated value file. The first line has column labels and the remaining lines are a list of gene conversion tracts, one tract per row.
       4. “*GC\_arr\_Transcripts\_List.tsv*”, lists the exons for all transcripts in *arr* coming from the “*arr* List” tab from “2019\_02\_02 Coordinates\_List.xlsx.”
       5. “*GC\_arr\_ML\_Exp.tsv*”, The file has 70 lines with the Maximum Likelihood estimates of the mean f or the probability of extending the gene conversion tract to an additional nucleotide in the ith gene arrangement in the jth region and the number of true gene conversion tracts in the ith arrangement in the jth region.
    4. Output Files
       1. “*GC\_arr\_Tracts\_Output\_MLpara.csv*” lists the intervals with gene conversion tracts. The file has four columns [**Region**, **Beg**, **End**, **Count**] where Region is the number of the region having at least one gene conversion tract covering the nucleotide; Beg is the first nucleotide of the region; End is the last nucleotide of the region; Count is the mean number of gene conversion tract coverage for the region.
       2. “*GC\_arr\_Transcript\_Tract\_Depth\_MLpara.csv*”, lists the mean gene conversion coverage for each exon of all transcripts. The file has six columns [**Transcript**, **Exon**, **Beg**, **End**, **Outlier**, **Mean\_GC**] where Transcript is the transcript name; Exon is the number of the exon; Beg is the first nucleotide of the exon; End is the last nucleotide of the exon; Outlier is the outlier status of the transcript where 1 is non-outlier and 2 is outlier; Mean\_GC, mean gene conversion coverage per nucleotide.
       3. “*GC\_arr\_Random\_Shuffle\_ML\_Data.csv*” a comma separated value file with the mean coverage for the outlier genes in each of 10,000 random permutations.
       4. “*GC\_arr\_Outlier\_Tract\_Depth\_MLpara.txt*”, a text file with the results of the random permutation test. A table is presented with three columns [Mean GC Bases, Variance GC Bases, Nucleotides] and two rows {Non-Outlier, Outlier]. The probability of mean GC coverage for outlier genes bases on the number of random permutations less than or equal to the observed values is given. Finally, the minimum and maximum value of 10,000 random permutations is presented. A file is generated for each arrangement [AR, ST, PP, CH, TL]. The random seed iseed is printed.

**Literature Cited**

Betran, E., J. Rozas, A. Navarro and A. Barbadilla, 1997 The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. Genetics 146**:** 89-99.

Nei, M., and T. Gojobori, 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution 3**:** 418-426.