Ejercicio de parseo de archivos FASTA

Pablo Vinuesa

2019-02-21

Contents

Presentación

Este código corresponde a unas prácticas escritas por Pablo Vinuesa para el manual de Sistemática Molecular y Bioinformática de la Facultad de Ciencias - UNAM, Abril 2018.

Para correr los ejercicios, asegúrate de tener el archivo recA__Bradyrhizobium__vinuesa.fna en el directorio actual de trabajo.

Búsqueda y descarga de secuencias en GenBank usando el sistema ENTREZ

El archivo recA__Bradyrhizobium__vinuesa.fna contiene secuencias del gen *recA* de bacterias del género *Bradyrhizobium* depositadas en GenBank por P. Vinuesa.

Este bloque muestra el comando usado para descargarlas. El comando debe pegarse en la ventana superior del sistema ENTREZ.

```
# pega esta sentencia en la ventana de captura para interrogar la base de datos
# de nucleótidos de NCBI mediante el sistema ENTREZ
'Bradyrhizobium[orgn] AND vinuesa[auth] AND recA[gene]'
```

Práctica de parseo de archivos FASTA descargados de NCBI mediante ENTREZ

Acceso a las secuencias

```
cd $HOME/intro2genomics
cp -r /home/vinuesa/cursos/intro2genomics/sesion1_parseo_fastas .
cd sesion1_parseo_fastas
```

Inspección y estadísticas básicas de las secuencias descargadas

1. ¿Cuántas secuencias hay en el archivo recA__Bradyrhizobium__vinuesa.fna?

```
grep -c '>' recA_Bradyrhizobium_vinuesa.fna
```

```
## 125
```

2. Veamos las 5 primeras lineas de cabeceras fasta usando **grep** y **head**

```
grep '>' recA_Bradyrhizobium_vinuesa.fna | head -5
```

```
## >gi|190612137|[B.liaoningense_strain_ViHaR5]
## >gi|190612135|[B.liaoningense_strain_ViHaR4]
## >gi|190612133|[B.liaoningense_strain_ViHaR3]
## >gi|190612131|[B.liaoningense_strain_ViHaR2]
## >gi|190612129|[B.liaoningense_strain_ViHaR1]
```

3. Cuenta el numero de generos y especies que contiene el archivo FASTA

```
grep '>' recA_Bradyrhizobium_vinuesa.fna | cut -d' ' -f3 | sort | uniq -c
```

```
##       1 >gi|190611979|[B.yuanmingense_strain_BuCeG2]
##       1 >gi|190611981|[B.yuanmingense_strain_BuCeG3]
##       1 >gi|190611983|[B.yuanmingense_strain_BuCeG4]
##       1 >gi|190611985|[B.sp._BuCeR1]
##       1 >gi|190611987|[B.sp._BuCeR2]
##       1 >gi|190611989|[B.yuanmingense_strain_BuCeR3]
##       1 >gi|190611991|[B.yuanmingense_strain_BuCeR4]
##       1 >gi|190611993|[B.yuanmingense_strain_BuCeR5]
##       1 >gi|190611995|[B.elkanii_strain_BuMiN1]
##       1 >gi|190611997|[B.elkanii_strain_BuMiN2]
##       1 >gi|190611999|[B.elkanii_strain_BuMiN3]
##       1 >gi|190612001|[B.elkanii_strain_BuMiN4]
##       1 >gi|190612003|[B.liaoningense_strain_BuMiN6]
##       1 >gi|190612005|[B.elkanii_strain_BuMiT1]
##       1 >gi|190612007|[B.sp._BuMiT10]
##       1 >gi|190612009|[B.liaoningense_strain_BuMiT3]
##       1 >gi|190612011|[B.liaoningense_strain_BuMiT4]
##       1 >gi|190612013|[B.liaoningense_strain_BuMiT5]
##       1 >gi|190612015|[B.elkanii_strain_BuMiT6]
##       1 >gi|190612017|[B.elkanii_strain_BuMiT7]
##       1 >gi|190612019|[B.elkanii_strain_BuMiT8]
##       1 >gi|190612021|[B.elkanii_strain_BuMiT9]
##       1 >gi|190612023|[B.elkanii_strain_BuNoG1]
##       1 >gi|190612025|[B.elkanii_strain_BuNoG4]
##       1 >gi|190612027|[B.sp._BuNoG5]
##       1 >gi|190612029|[B.elkanii_strain_BuNoR1]
##       1 >gi|190612031|[B.elkanii_strain_BuNoR2]
##       1 >gi|190612033|[B.elkanii_strain_BuNoR3]
##       1 >gi|190612035|[B.elkanii_strain_BuNoR4]
##       1 >gi|190612037|[B.yuanmingense_strain_InBu02]
##       1 >gi|190612039|[B.yuanmingense_strain_InIn01]
##       1 >gi|190612041|[B.yuanmingense_strain_InIn02]
##       1 >gi|190612043|[B.yuanmingense_strain_InIn03]
##       1 >gi|190612045|[B.yuanmingense_strain_InIn04]
##       1 >gi|190612047|[B.yuanmingense_strain_InIn05]
##       1 >gi|190612049|[B.yuanmingense_strain_InIn08]
##       1 >gi|190612051|[B.yuanmingense_strain_InIn09]
##       1 >gi|190612053|[B.yuanmingense_strain_InIn10]
##       1 >gi|190612055|[B.yuanmingense_strain_InJa01]
##       1 >gi|190612057|[B.yuanmingense_strain_InJa02]
##       1 >gi|190612059|[B.yuanmingense_strain_InJa03]
##       1 >gi|190612061|[B.yuanmingense_strain_InJa04]
##        1 >gi|190612063|[B.yuanmingense_strain_InJa05]
##        1 >gi|190612065|[B.yuanmingense_strain_InJa06]
##        1 >gi|190612067|[B.yuanmingense_strain_InJa07]
##        1 >gi|190612069|[B.yuanmingense_strain_InJa08]
##        1 >gi|190612071|[B.yuanmingense_strain_InJa09]
##        1 >gi|190612073|[B.yuanmingense_strain_InKo01]
##        1 >gi|190612075|[B.yuanmingense_strain_InKo02]
##        1 >gi|190612077|[B.yuanmingense_strain_InRo02]
##        1 >gi|190612079|[B.japonicum_strain_NeMa01]
##        1 >gi|190612081|[B.japonicum_strain_NeMa02]
##        1 >gi|190612083|[B.japonicum_strain_NeMa10]
##        1 >gi|190612085|[B.japonicum_strain_NeMa11]
##        1 >gi|190612087|[B.japonicum_strain_NeMa12]
##        1 >gi|190612089|[B.japonicum_strain_NeMa16]
##        1 >gi|190612091|[B.japonicum_strain_NeRa01]
##        1 >gi|190612093|[B.japonicum_strain_NeRa02]
##        1 >gi|190612095|[B.japonicum_strain_NeRa03]
##        1 >gi|190612097|[B.japonicum_strain_NeRa04]
##        1 >gi|190612099|[B.japonicum_strain_NeRa05]
##        1 >gi|190612101|[B.japonicum_strain_NeRa06]
##        1 >gi|190612103|[B.japonicum_strain_NeRa07]
##        1 >gi|190612105|[B.japonicum_strain_NeRa08]
##        1 >gi|190612107|[B.japonicum_strain_NeRa11]
##        1 >gi|190612109|[B.japonicum_strain_NeRa12]
##        1 >gi|190612111|[B.japonicum_strain_NeRa14]
##        1 >gi|190612113|[B.japonicum_strain_NeRa15]
##        1 >gi|190612115|[B.japonicum_strain_NeRa16]
##        1 >gi|190612117|[B.liaoningense_strain_ViHaG3]
##        1 >gi|190612119|[B.yuanmingense_strain_ViHaG4]
##        1 >gi|190612121|[B.yuanmingense_strain_ViHaG5]
##        1 >gi|190612123|[B.liaoningense_strain_ViHaG6]
##        1 >gi|190612125|[B.liaoningense_strain_ViHaG7]
##        1 >gi|190612127|[B.liaoningense_strain_ViHaG8]
##        1 >gi|190612129|[B.liaoningense_strain_ViHaR1]
##        1 >gi|190612131|[B.liaoningense_strain_ViHaR2]
##        1 >gi|190612133|[B.liaoningense_strain_ViHaR3]
##        1 >gi|190612135|[B.liaoningense_strain_ViHaR4]
##        1 >gi|190612137|[B.liaoningense_strain_ViHaR5]
##        1 >gi|50982176|[B.genosp._alpha_bv._genistearum_strain_BC-C1]
##        1 >gi|50982178|[B.canariense_bv._genistearum_strain_BC-C2]
##        1 >gi|50982180|[B.canariense_bv._genistearum_strain_BC-P5]
##        1 >gi|50982182|[B.genosp._beta_strain_BC-P6]
##        1 >gi|50982184|[B.japonicum_bv._genistearum_strain_BC-P14]
##        1 >gi|50982186|[B.canariense_bv._genistearum_strain_BC-P22]
##        1 >gi|50982188|[B.canariense_bv._genistearum_strain_BC-MAM1]
##        1 >gi|50982190|[B.canariense_bv._genistearum_strain_BC-MAM5]
##        1 >gi|50982192|[B.canariense_bv._genistearum_strain_BES-1]
##        1 >gi|50982194|[B.canariense_bv._genistearum_strain_BES-2]
##        1 >gi|50982196|[B.canariense_bv._genistearum_strain_BC0-1]
##        1 >gi|50982198|[B.genosp._beta_strain_BRE-1]
##        1 >gi|50982200|[B.canariense_bv._genistearum_strain_BRE-4]
##        1 >gi|50982202|[B.canariense_bv._genistearum_strain_BTA-1]
##        1 >gi|50982204|[B.genosp._beta_strain_BC-MK6]
##        1 >gi|50982206|[B.japonicum_bv._glycinearum_strain_DSMZ30131]
##       1 >gi|50982208|[B.japonicum_bv._glycinearum_strain_X3-1]
##       1 >gi|50982210|[B.japonicum_bv._glycinearum_strain_X6-9]
##       1 >gi|50982212|[B.japonicum_bv._genistearum_strain_BGA-1]
##       1 >gi|50982214|[B.japonicum_bv._genistearum_strain_BLup-MR1]
##       1 >gi|50982216|[B.japonicum_bv._genistearum_strain_FN13]
##       1 >gi|50982218|[B.sp._CICS70]
##       1 >gi|50982220|[B.japonicum_bv._glycinearum_strain_USDA122]
##       1 >gi|50982222|[B.japonicum_bv._glycinearum_strain_Nep1]
##       1 >gi|50982224|[B.liaoningense_bv._glycinearum_strain_LMG18230]
##       1 >gi|50982226|[B.yuanmingense_strain_TAL760]
##       1 >gi|50982228|[B.yuanmingense_strain_CCBAU_10071]
##       1 >gi|50982230|[B.genosp._alpha_strain_CIAT3101]
##       1 >gi|50982232|[B.elkanii_strain_USDA76]
##       1 >gi|50982234|[B.elkanii_strain_USDA94]
##       1 >gi|50982236|[B.sp._BTAi1]
##       1 >gi|50982238|[B.sp._IRBG127]
##       1 >gi|50982240|[B.sp._IRBG231]
##       1 >gi|50982242|[B.yuanmingense_strain_LMTR28]
##       1 >gi|50982244|[B.liaoningense_strain_Spr3-7]
##       1 >gi|50982246|[B.elkanii_strain_USDA46]
##       1 >gi|50982248|[B.canariense_strain_ISLU16]
##       1 >gi|52550802|[B.canariense_strain_BC-P24]
##       1 >gi|52550804|[B.canariense_strain_BC-MAM2]
##       1 >gi|52550806|[B.canariense_strain_BC-MAM6]
##       1 >gi|52550808|[B.canariense_strain_BC-MAM8]
##       1 >gi|52550810|[B.canariense_strain_BC-MAM9]
##       1 >gi|52550812|[B.canariense_strain_BC-MAM11]
##       1 >gi|52550814|[B.canariense_strain_BC-MAM12]
##       1 >gi|52550816|[B.genosp._beta_strain_BC-MK1]
```

4. Imprime una lista ordenada de mayor a menor, del numero de especies que contiene el archivo FASTA

```
grep '>' recA_Bradyrhizobium_vinuesa.fna | cut -d' ' -f2,3 | sort | uniq -c | sort -nrk1
```

```
##       1 >gi|52550816|[B.genosp._beta_strain_BC-MK1]
##       1 >gi|52550814|[B.canariense_strain_BC-MAM12]
##       1 >gi|52550812|[B.canariense_strain_BC-MAM11]
##       1 >gi|52550810|[B.canariense_strain_BC-MAM9]
##       1 >gi|52550808|[B.canariense_strain_BC-MAM8]
##       1 >gi|52550806|[B.canariense_strain_BC-MAM6]
##       1 >gi|52550804|[B.canariense_strain_BC-MAM2]
##       1 >gi|52550802|[B.canariense_strain_BC-P24]
##       1 >gi|50982248|[B.canariense_strain_ISLU16]
##       1 >gi|50982246|[B.elkanii_strain_USDA46]
##       1 >gi|50982244|[B.liaoningense_strain_Spr3-7]
##       1 >gi|50982242|[B.yuanmingense_strain_LMTR28]
##       1 >gi|50982240|[B.sp._IRBG231]
##       1 >gi|50982238|[B.sp._IRBG127]
##       1 >gi|50982236|[B.sp._BTAi1]
##       1 >gi|50982234|[B.elkanii_strain_USDA94]
##       1 >gi|50982232|[B.elkanii_strain_USDA76]
##       1 >gi|50982230|[B.genosp._alpha_strain_CIAT3101]
##       1 >gi|50982228|[B.yuanmingense_strain_CCBAU_10071]
##       1 >gi|50982226|[B.yuanmingense_strain_TAL760]
##       1 >gi|50982224|[B.liaoningense_bv._glycinearum_strain_LMG18230]
##       1 >gi|50982222|[B.japonicum_bv._glycinearum_strain_Nep1]
##       1 >gi|50982220|[B.japonicum_bv._glycinearum_strain_USDA122]
##       1 >gi|50982218|[B.sp._CICS70]
##       1 >gi|50982216|[B.japonicum_bv._genistearum_strain_FN13]
##       1 >gi|50982214|[B.japonicum_bv._genistearum_strain_BLup-MR1]
##       1 >gi|50982212|[B.japonicum_bv._genistearum_strain_BGA-1]
##       1 >gi|50982210|[B.japonicum_bv._glycinearum_strain_X6-9]
##       1 >gi|50982208|[B.japonicum_bv._glycinearum_strain_X3-1]
##       1 >gi|50982206|[B.japonicum_bv._glycinearum_strain_DSMZ30131]
##       1 >gi|50982204|[B.genosp._beta_strain_BC-MK6]
##       1 >gi|50982202|[B.canariense_bv._genistearum_strain_BTA-1]
##       1 >gi|50982200|[B.canariense_bv._genistearum_strain_BRE-4]
##       1 >gi|50982198|[B.genosp._beta_strain_BRE-1]
##       1 >gi|50982196|[B.canariense_bv._genistearum_strain_BC0-1]
##       1 >gi|50982194|[B.canariense_bv._genistearum_strain_BES-2]
##       1 >gi|50982192|[B.canariense_bv._genistearum_strain_BES-1]
##       1 >gi|50982190|[B.canariense_bv._genistearum_strain_BC-MAM5]
##       1 >gi|50982188|[B.canariense_bv._genistearum_strain_BC-MAM1]
##       1 >gi|50982186|[B.canariense_bv._genistearum_strain_BC-P22]
##       1 >gi|50982184|[B.japonicum_bv._genistearum_strain_BC-P14]
##       1 >gi|50982182|[B.genosp._beta_strain_BC-P6]
##       1 >gi|50982180|[B.canariense_bv._genistearum_strain_BC-P5]
##       1 >gi|50982178|[B.canariense_bv._genistearum_strain_BC-C2]
##       1 >gi|50982176|[B.genosp._alpha_bv._genistearum_strain_BC-C1]
##       1 >gi|190612137|[B.liaoningense_strain_ViHaR5]
##       1 >gi|190612135|[B.liaoningense_strain_ViHaR4]
##       1 >gi|190612133|[B.liaoningense_strain_ViHaR3]
##       1 >gi|190612131|[B.liaoningense_strain_ViHaR2]
##       1 >gi|190612129|[B.liaoningense_strain_ViHaR1]
##       1 >gi|190612127|[B.liaoningense_strain_ViHaG8]
##       1 >gi|190612125|[B.liaoningense_strain_ViHaG7]
##       1 >gi|190612123|[B.liaoningense_strain_ViHaG6]
##       1 >gi|190612121|[B.yuanmingense_strain_ViHaG5]
##       1 >gi|190612119|[B.yuanmingense_strain_ViHaG4]
##       1 >gi|190612117|[B.liaoningense_strain_ViHaG3]
##       1 >gi|190612115|[B.japonicum_strain_NeRa16]
##       1 >gi|190612113|[B.japonicum_strain_NeRa15]
##       1 >gi|190612111|[B.japonicum_strain_NeRa14]
##       1 >gi|190612109|[B.japonicum_strain_NeRa12]
##       1 >gi|190612107|[B.japonicum_strain_NeRa11]
##       1 >gi|190612105|[B.japonicum_strain_NeRa08]
##       1 >gi|190612103|[B.japonicum_strain_NeRa07]
##       1 >gi|190612101|[B.japonicum_strain_NeRa06]
##       1 >gi|190612099|[B.japonicum_strain_NeRa05]
##       1 >gi|190612097|[B.japonicum_strain_NeRa04]
##       1 >gi|190612095|[B.japonicum_strain_NeRa03]
##       1 >gi|190612093|[B.japonicum_strain_NeRa02]
##       1 >gi|190612091|[B.japonicum_strain_NeRa01]
##       1 >gi|190612089|[B.japonicum_strain_NeMa16]
##       1 >gi|190612087|[B.japonicum_strain_NeMa12]
##       1 >gi|190612085|[B.japonicum_strain_NeMa11]
##       1 >gi|190612083|[B.japonicum_strain_NeMa10]
##       1 >gi|190612081|[B.japonicum_strain_NeMa02]
##       1 >gi|190612079|[B.japonicum_strain_NeMa01]
##       1 >gi|190612077|[B.yuanmingense_strain_InRo02]
##       1 >gi|190612075|[B.yuanmingense_strain_InKo02]
##       1 >gi|190612073|[B.yuanmingense_strain_InKo01]
##       1 >gi|190612071|[B.yuanmingense_strain_InJa09]
##       1 >gi|190612069|[B.yuanmingense_strain_InJa08]
##       1 >gi|190612067|[B.yuanmingense_strain_InJa07]
##       1 >gi|190612065|[B.yuanmingense_strain_InJa06]
##       1 >gi|190612063|[B.yuanmingense_strain_InJa05]
##       1 >gi|190612061|[B.yuanmingense_strain_InJa04]
##       1 >gi|190612059|[B.yuanmingense_strain_InJa03]
##       1 >gi|190612057|[B.yuanmingense_strain_InJa02]
##       1 >gi|190612055|[B.yuanmingense_strain_InJa01]
##       1 >gi|190612053|[B.yuanmingense_strain_InIn10]
##       1 >gi|190612051|[B.yuanmingense_strain_InIn09]
##       1 >gi|190612049|[B.yuanmingense_strain_InIn08]
##       1 >gi|190612047|[B.yuanmingense_strain_InIn05]
##       1 >gi|190612045|[B.yuanmingense_strain_InIn04]
##       1 >gi|190612043|[B.yuanmingense_strain_InIn03]
##       1 >gi|190612041|[B.yuanmingense_strain_InIn02]
##       1 >gi|190612039|[B.yuanmingense_strain_InIn01]
##       1 >gi|190612037|[B.yuanmingense_strain_InBu02]
##       1 >gi|190612035|[B.elkanii_strain_BuNoR4]
##       1 >gi|190612033|[B.elkanii_strain_BuNoR3]
##       1 >gi|190612031|[B.elkanii_strain_BuNoR2]
##       1 >gi|190612029|[B.elkanii_strain_BuNoR1]
##       1 >gi|190612027|[B.sp._BuNoG5]
##       1 >gi|190612025|[B.elkanii_strain_BuNoG4]
##       1 >gi|190612023|[B.elkanii_strain_BuNoG1]
##       1 >gi|190612021|[B.elkanii_strain_BuMiT9]
##       1 >gi|190612019|[B.elkanii_strain_BuMiT8]
##       1 >gi|190612017|[B.elkanii_strain_BuMiT7]
##       1 >gi|190612015|[B.elkanii_strain_BuMiT6]
##       1 >gi|190612013|[B.liaoningense_strain_BuMiT5]
##       1 >gi|190612011|[B.liaoningense_strain_BuMiT4]
##       1 >gi|190612009|[B.liaoningense_strain_BuMiT3]
##       1 >gi|190612007|[B.sp._BuMiT10]
##       1 >gi|190612005|[B.elkanii_strain_BuMiT1]
##       1 >gi|190612003|[B.liaoningense_strain_BuMiN6]
##       1 >gi|190612001|[B.elkanii_strain_BuMiN4]
##       1 >gi|190611999|[B.elkanii_strain_BuMiN3]
##       1 >gi|190611997|[B.elkanii_strain_BuMiN2]
##       1 >gi|190611995|[B.elkanii_strain_BuMiN1]
##       1 >gi|190611993|[B.yuanmingense_strain_BuCeR5]
##       1 >gi|190611991|[B.yuanmingense_strain_BuCeR4]
##       1 >gi|190611989|[B.yuanmingense_strain_BuCeR3]
##       1 >gi|190611987|[B.sp._BuCeR2]
##       1 >gi|190611985|[B.sp._BuCeR1]
##       1 >gi|190611983|[B.yuanmingense_strain_BuCeG4]
##       1 >gi|190611981|[B.yuanmingense_strain_BuCeG3]
##       1 >gi|190611979|[B.yuanmingense_strain_BuCeG2]
```

Edición de las cabeceras FASTA mediante herramientas de filtrado de UNIX

5. Exploremos todas las cabeceras FASTA del archivo recA_Bradyrhizobium_vinuesa.fna usando **grep**

```
# grep '>' recA_Bradyrhizobium_vinuesa.fna | less # para verlas por página
grep '>' recA_Bradyrhizobium_vinuesa.fna  | head # para no hacer muy extensa la salida
```

```
## >gi|190612137|[B.liaoningense_strain_ViHaR5]
## >gi|190612135|[B.liaoningense_strain_ViHaR4]
## >gi|190612133|[B.liaoningense_strain_ViHaR3]
## >gi|190612131|[B.liaoningense_strain_ViHaR2]
## >gi|190612129|[B.liaoningense_strain_ViHaR1]
## >gi|190612127|[B.liaoningense_strain_ViHaG8]
## >gi|190612125|[B.liaoningense_strain_ViHaG7]
## >gi|190612123|[B.liaoningense_strain_ViHaG6]
## >gi|190612121|[B.yuanmingense_strain_ViHaG5]
## >gi|190612119|[B.yuanmingense_strain_ViHaG4]
```

6. simplifiquemos las cabeceras FASTA usando el comando **sed** (stream editor)

El objetivo es eliminar redundancia y los campos gb|no.de.acceso, así como todos los caracteres '(, ; :)' que impedirían el despliegue de un árbol filogenético, al tratarse de caracteres reservados del formato NEWICK. Dejar solo el numero GI, así como el género, especie y cepa indicados entre corchetes.

Es decir vamos a: - reducir Bradyrhizobium a 'B.' - eliminar 'RNA poly . . . ' y reemplazarlo por ']' - eliminar 'genosp.' - sustituir espacios por guiones bajos

Noten el uso de expresiones regulares como '.*'y'[[:space:]]'

```
sed 's/ Bra/ [Bra/; s/|gb.*| /|/; s/Bradyrhizobium /B./; s/genosp\. //; s/ RNA.*/]/; s/[[:space:]]/_/g;
```

```
## >gi|190612137|[B.liaoningense_strain_ViHaR5]
## >gi|190612135|[B.liaoningense_strain_ViHaR4]
## >gi|190612133|[B.liaoningense_strain_ViHaR3]
## >gi|190612131|[B.liaoningense_strain_ViHaR2]
## >gi|190612129|[B.liaoningense_strain_ViHaR1]
```

8. Cuando estamos satisfechos con el resultado, guardamos la salida del comando en un archivo usando '>' para redirigir el flujo de STDOUT a un archivo de texto

```
sed 's/ recom.*cds/]/; s/ Bra/ [Bra/; s/|gb.*| /|/; s/Bradyrhizobium /B. /; s/genosp\. //; s/ RNA.*/\]/
```

Generación automática de archivos FASTA especie-específicos (avanzado)

9. Convertir archivos FASTA a formato "FASTAB" usando **perl** 1-liners.

Vamos a transformar los FASTAS de tal manera que las secuencias queden en la misma línea que su cabecera, separada de ésta por un tabulador. Esto puede ser muy útil para filtrar el archivo resultante con grep. Veamos un ejemplo:

```
perl -pe 'unless(/^>/){s/\n//g}; if(/>/){s/\n/\t/g}; s/>/\n>/' recA_Bradyrhizobium_vinuesa.fnaed | head
```

```
##
## >gi|190612137|[B.liaoningense_strain_ViHaR5] ATGAAGCTCGGCAAGAACGACCGGTCCATGGACATCGAGGCGGTGTCCTCCGGCT
## >gi|190612135|[B.liaoningense_strain_ViHaR4] ATGAAGCTCGGCAAGAACGACCGGTCCATGGACATCGAGGCGGTGTCCTCCGGCT
## >gi|190612133|[B.liaoningense_strain_ViHaR3] ATGAAGCTCGGCAAGAACGACCGGTCCATGGACATCGAGGCGGTGTCCTCCGGCT
## >gi|190612131|[B.liaoningense_strain_ViHaR2] ATGAAGCTCGGCAAGAACGACCGGTCCATGGACATCGAGGCGGTGTCCTCCGGCT
## >gi|190612129|[B.liaoningense_strain_ViHaR1] ATGAAGCTCGGCAAGAACGACCGGTCCATGGACATCGAGGCGGTGTCCTCCGGCT
## >gi|190612127|[B.liaoningense_strain_ViHaG8] ATGAAGCTCGGCAAGAACGACCGGTCCATGGACATCGAGGCGGTGTCCTCCGGCT
## >gi|190612125|[B.liaoningense_strain_ViHaG7] ATGAAGCTCGGCAAGAACGACCGGTCCATGGACATCGAGGCGGTGTCCTCCGGCT
## >gi|190612123|[B.liaoningense_strain_ViHaG6] ATGAAGCTCGGCAAGAACGACCGGTCCATGGACATCGAGGCGGTGTCCTCCGGCT
## >gi|190612121|[B.yuanmingense_strain_ViHaG5] ATGAAGCTCGGCAAGAACGACCGCTCCATGGACATCGAGGCGGTGTCCTCCGGCT
```

```
perl -pe 'unless(/^>/){s/\n//g}; if(/>/){s/\n/\t/g}; s/>/\n>/' recA_Bradyrhizobium_vinuesa.fnaed > recA_
```

10. Filtrar el archivo fnaedtab generado en 9 para obtener solo las secuencias de B.__yuanmingense del mismo, guardarlo en un archivo y convertirlo de nuevo a formato FASTA.

```
grep yuanmingense recA_Bradyrhizobium_vinuesa.fnaedtab | head -5
```

```
## >gi|190612121|[B.yuanmingense_strain_ViHaG5] ATGAAGCTCGGCAAGAACGACCGCTCCATGGACATCGAGGCGGTGTCCTCCGGCT
## >gi|190612119|[B.yuanmingense_strain_ViHaG4] ATGAAGCTCGGCAAGAACGACCGCTCCATGGACATCGAGGCGGTGTCCTCCGGCT
## >gi|190612077|[B.yuanmingense_strain_InRo02] ATGAAGCTCGGCAAGAACGACCGCTCCATGGACATCGAGGCGGTGTCCTCCGGCT
## >gi|190612075|[B.yuanmingense_strain_InKo02] ATGAAGCTCGGCAAGAACGATCGCTCCATGGACATCGAGGCGGTCTCCTCCGGCT
## >gi|190612073|[B.yuanmingense_strain_InKo01] ATGAAGCTCGGCAAGAACGATCGCTCCATGGACATCGAGGCGGTGTCCTCCGGCT
```

```
grep yuanmingense recA_Bradyrhizobium_vinuesa.fnaedtab > recA_Byuanmingense.fnaedtab
```

11. Estas dos lineas no contienen nada nuevo en cuanto a sintaxis. Simplemente llamamos a perl para sustituir los tabuladores por saltos de linea y asi reconstituir el FASTA.

```
perl -pe 'if(/>/){s/\t/\n/}' recA_Byuanmingense.fnaedtab | head -5
```

```
## >gi|190612121|[B.yuanmingense_strain_ViHaG5]
## ATGAAGCTCGGCAAGAACGACCGCTCCATGGACATCGAGGCGGTGTCCTCCGGCTCGCTCGGGCTCGATATCGCGCTCGGCATCGGCGGCTTGCCCAAGG
## >gi|190612119|[B.yuanmingense_strain_ViHaG4]
## ATGAAGCTCGGCAAGAACGACCGCTCCATGGACATCGAGGCGGTGTCCTCCGGCTCGCTCGGGCTCGATATCGCGCTCGGCATCGGCGGCTTGCCCAAGG
## >gi|190612077|[B.yuanmingense_strain_InRo02]
```

```
perl -pe 'if(/>/){s/\t/\n/}' recA_Byuanmingense.fnaedtab > recA_Byuanmingense.fna
```

12. Llamar a un bucle for de shell para generar archivos fastab para todas las especies

```
for sp in $(grep '>' recA_Bradyrhizobium_vinuesa.fnaed | cut -d_ -f2); do
   grep "$sp" recA_Bradyrhizobium_vinuesa.fnaedtab > recA_B${sp}.fnaedtab
done
```

13. Veamos el resultado

```
ls *fnaedtab
```

```
## recA_Balpha.fnaedtab
## recA_Bbeta.fnaedtab
## recA_BBTAi1].fnaedtab
## recA_BBuCeR1].fnaedtab
## recA_BBuCeR2].fnaedtab
## recA_BBuMiT10].fnaedtab
## recA_BBuNoG5].fnaedtab
## recA_Bbv..fnaedtab
## recA_Bcanariense.fnaedtab
## recA_BCICS70].fnaedtab
## recA_Belkanii.fnaedtab
## recA_Bgenosp.fnaedtab
## recA_BIRBG127].fnaedtab
## recA_BIRBG231].fnaedtab
## recA_Bjaponicum.fnaedtab
## recA_Bliaoningense.fnaedtab
## recA_Bradyrhizobium_vinuesa.fnaedtab
## recA_Bsp.fnaedtab
## recA_Bstrain.fnaedtab
## recA_Byuanmingense.fnaedtab
```

```
head -5 recA_Bjaponicum.fnaedtab
```

```
## >gi|190612115|[B.japonicum_strain_NeRa16]    ATGAAGCTCGGCAAGAACGACCGGTCGATGGATGTCGAGGCGGTGTCCTCGGGTT
## >gi|190612113|[B.japonicum_strain_NeRa15]    ATGAAGCTCGGCAAGAACGACCGGTCGATGGATGTCGAGGCGGTGTCCTCCGGTT
## >gi|190612111|[B.japonicum_strain_NeRa14]    ATGAAGCTCGGCAAGAACGACCGGTCGATGGATGTCGAGGCGGTGTCCTCCGGTT
## >gi|190612109|[B.japonicum_strain_NeRa12]    ATGAAGCTCGGCAAGAACGACCGGTCGATGGATGTCGAGGCGGTGTCCTCCGGTT
## >gi|190612107|[B.japonicum_strain_NeRa11]    ATGAAGCTCGGCAAGAACGACCGGTCGATGGATGTCGAGGCGGTGTCCTCCGGTT
```

14. Finalmente convertimos todos los archivos fnatabed a FASTA con el siguiente bucle for:

```
for file in *fnaedtab; do perl -pe 'if(/>/){s/\t/\n/}' $file > ${file%.*}.fna; done
```

15. Visualizemos las cabeceras de dos archivos FASTA especie-específicos

```
grep '>' recA_Bjaponicum.fna | head -5
```

```
## >gi|190612115|[B.japonicum_strain_NeRa16]
## >gi|190612113|[B.japonicum_strain_NeRa15]
## >gi|190612111|[B.japonicum_strain_NeRa14]
## >gi|190612109|[B.japonicum_strain_NeRa12]
## >gi|190612107|[B.japonicum_strain_NeRa11]
```

16. y confirmemos que son fastas regulares

```
head -6 recA_Bjaponicum.fna
```

```
## >gi|190612115|[B.japonicum_strain_NeRa16]
## ATGAAGCTCGGCAAGAACGACCGGTCGATGGATGTCGAGGCGGTGTCCTCGGGTTCTCTCGGGCTCGACATTGCACTGGGGATCGGCGGTCTGCCCAAGG
## >gi|190612113|[B.japonicum_strain_NeRa15]
## ATGAAGCTCGGCAAGAACGACCGGTCGATGGATGTCGAGGCGGTGTCCTCCGGTTCTCTCGGGCTCGACATTGCACTGGGGATCGGCGGTCTGCCCAAGG
## >gi|190612111|[B.japonicum_strain_NeRa14]
## ATGAAGCTCGGCAAGAACGACCGGTCGATGGATGTCGAGGCGGTGTCCTCCGGTTCTCTCGGGCTCGACATTGCGCTGGGGATCGGCGGTCTGCCCAAGG
```