tidyGenR
tidy multilocus amplicon genotypes in R

III Congreso & XIV Jornadas de Usuarios de R
Sevilla 2024

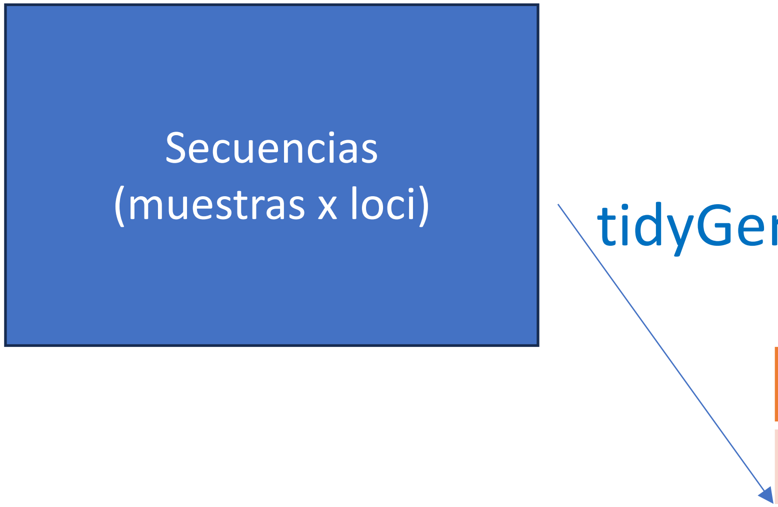
Miguel Camacho Sánchez

Estación Biológica de Doñana, Sevilla

*miguelcamachosanchez@gmail.com

IIIRqueR

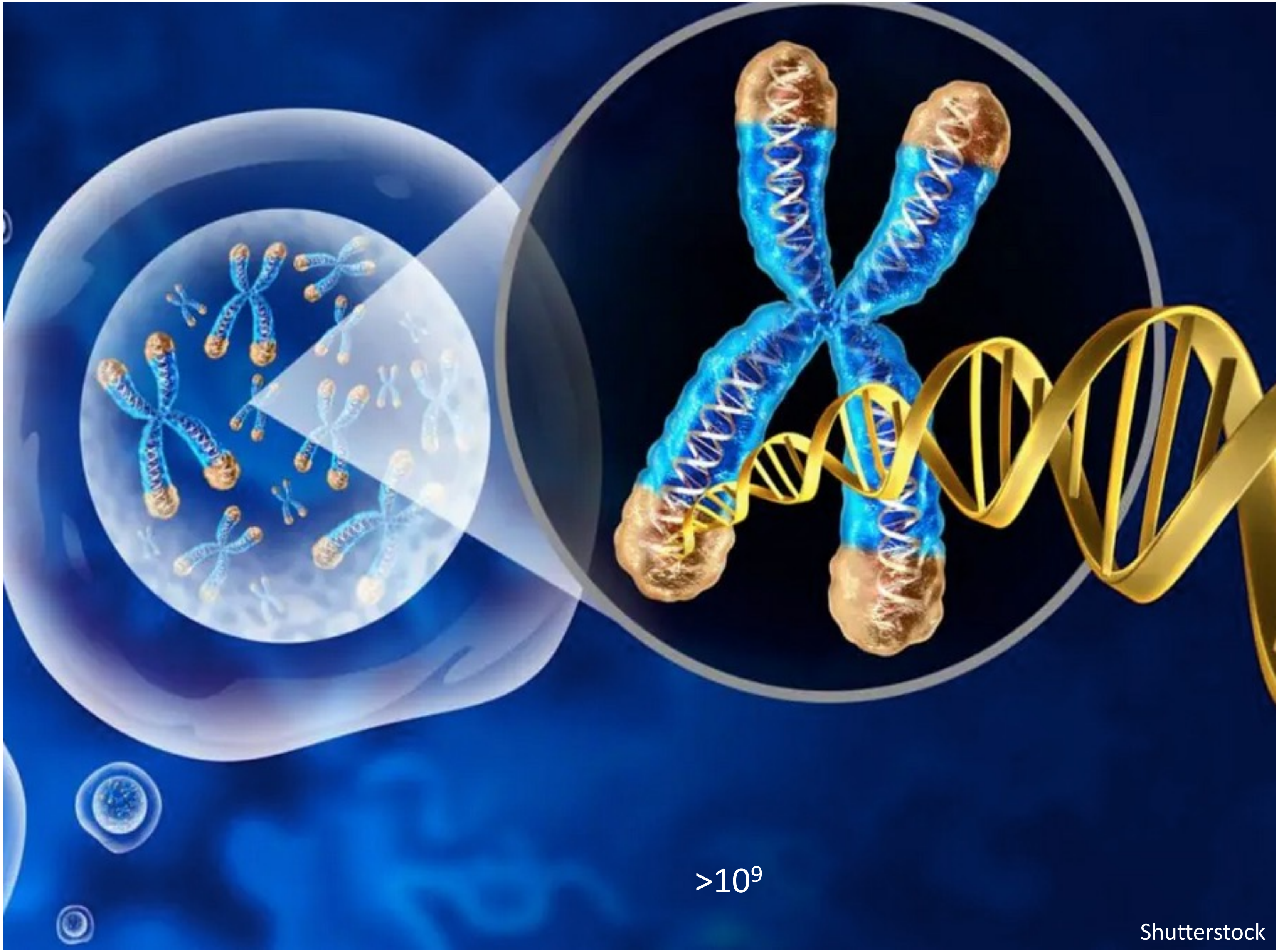
Genotipos a partir de lecturas de secuenciación masiva de múltiples loci


Secuencias
(muestras x loci)

tidyGenR

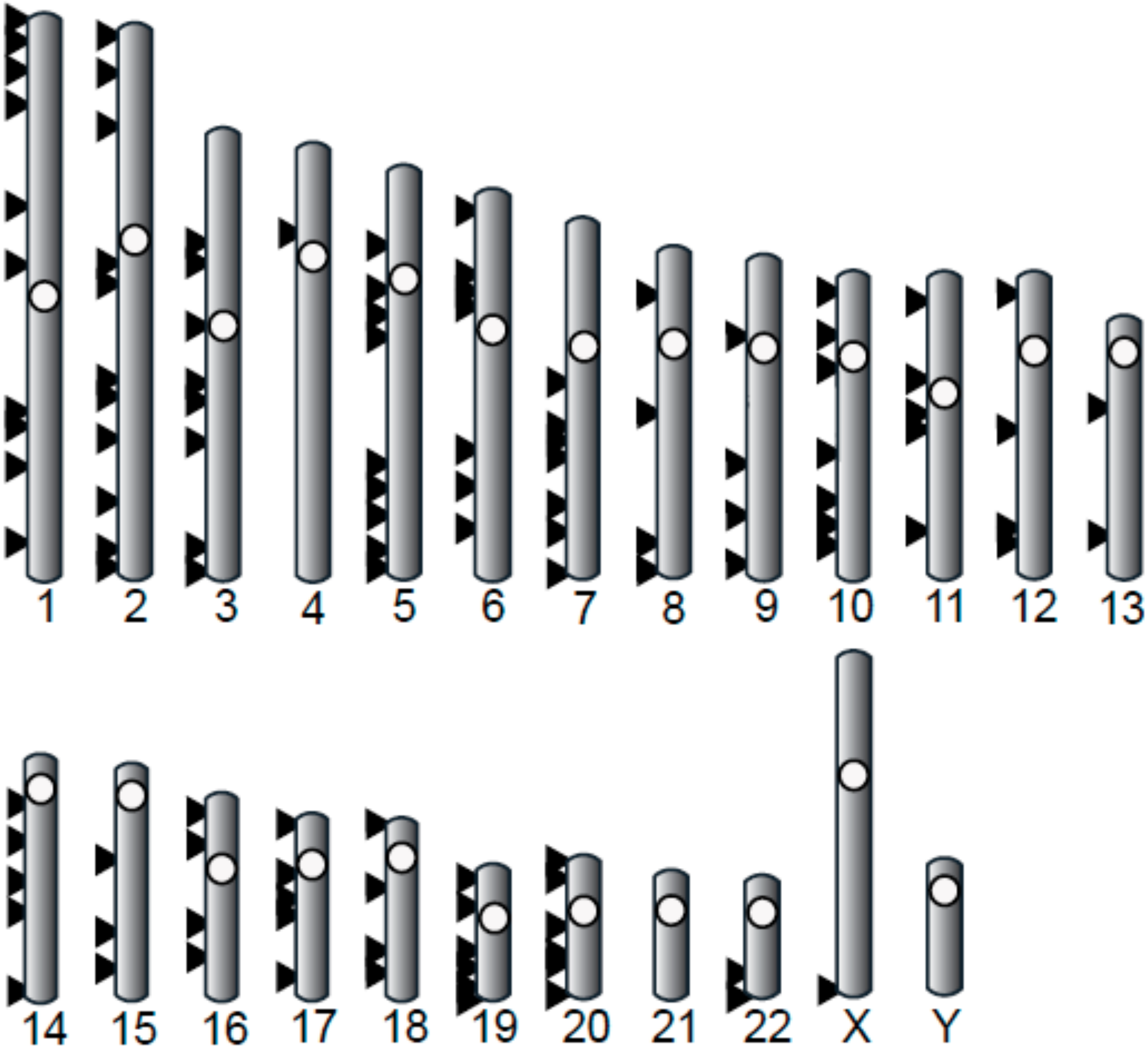
	locus_a	locus_b	locus_c
muestra_1	a/a	a/-	a/b
muestra_2	-/-	a/a	b/c
muestra_3	a/a	a/a	b/b

https://www.comunidad.madrid

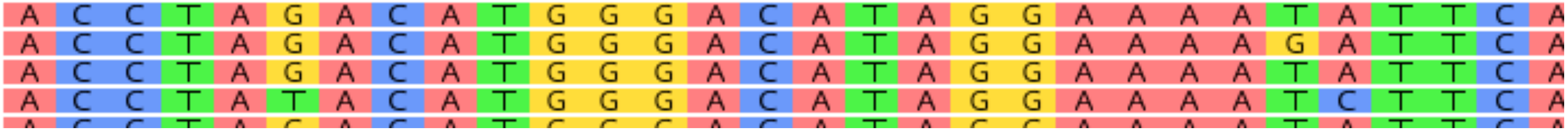


Shutterstock

IIIRqueR

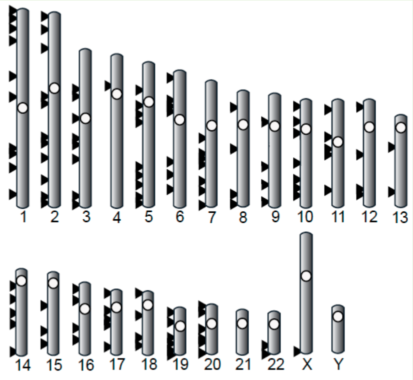


Igea et al. 2010

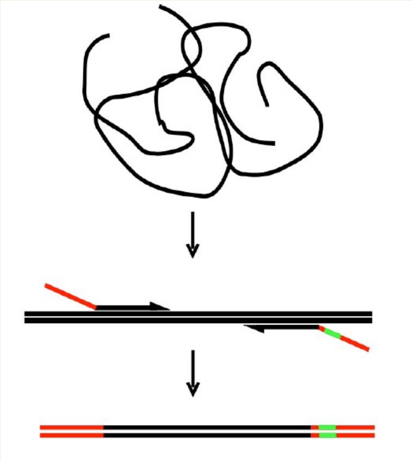


```
A C C T A G A C A T G G G A C A T A G G A A A A T A T T C A
A C C T A G A C A T G G G A C A T A G G A A A A G A T T C A
A C C T A G A C A T G G G A C A T A G G A A A A T A T T C A
A C C T A T A C A T G G G A C A T A G G A A A A T C T T C A
A C C T A G A C A T G G G A C A T A G G A A A A T A T T C A
```

Selección de marcadores

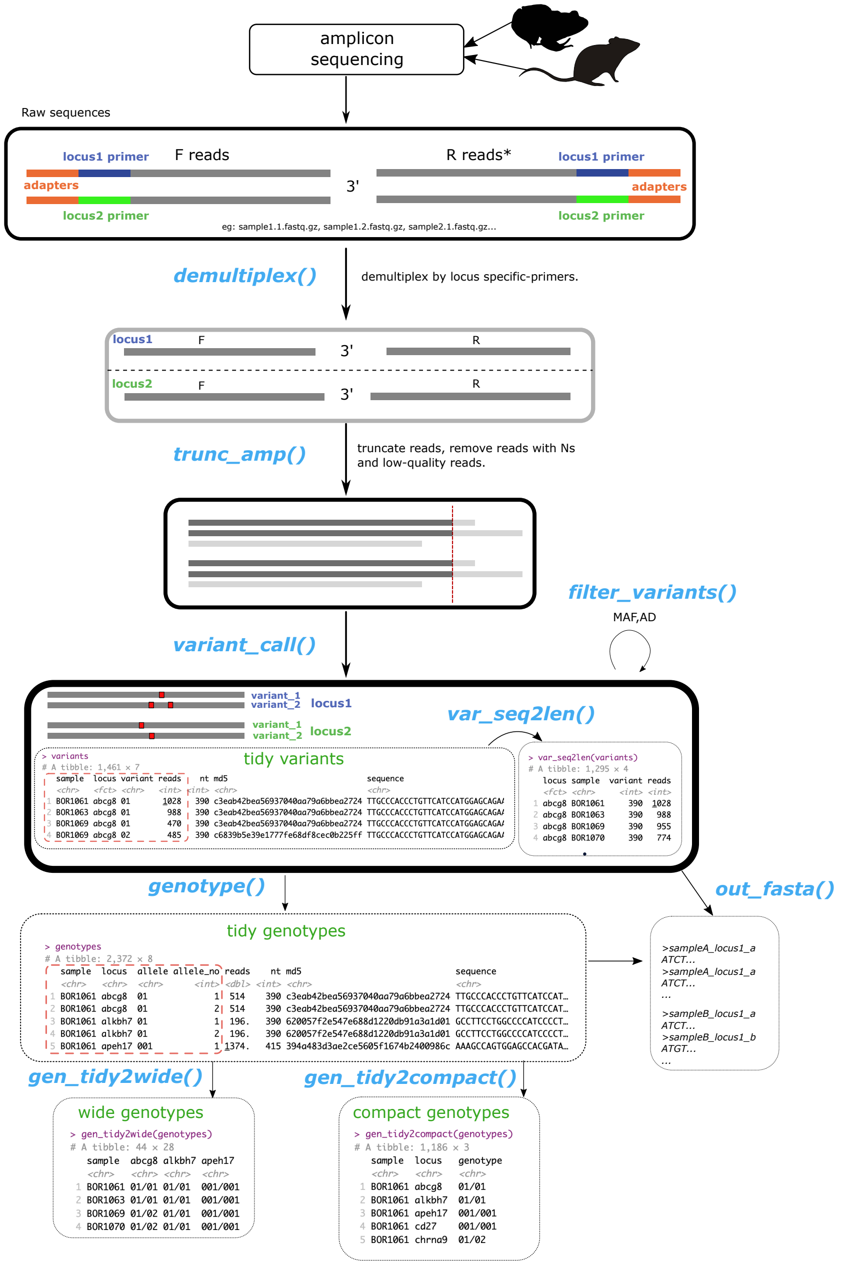


PCRs: Librerías de amplicones



Secuenciación masiva

Gb de secuencias → tidyGenR → Genotipos



```
> variants
# A tibble: 1,461 × 7
  sample  locus variant reads    nt md5                              sequence
  <chr>   <fct> <chr>   <int> <int> <chr>                            <chr>
1 BOR1061 abcg8 01       1028   390 c3eab42bea56937040aa79a6bbea2724 TTGCCCACCCTGTTCATCCATGGAGCAGA/
2 BOR1063 abcg8 01        988   390 c3eab42bea56937040aa79a6bbea2724 TTGCCCACCCTGTTCATCCATGGAGCAGA/
3 BOR1069 abcg8 01        470   390 c3eab42bea56937040aa79a6bbea2724 TTGCCCACCCTGTTCATCCATGGAGCAGA/
4 BOR1069 abcg8 02        485   390 c6839b5e39e1777fe68df8cec0b225ff TTGCCCACCCTGTTCATCCATGGAGCAGA/
```

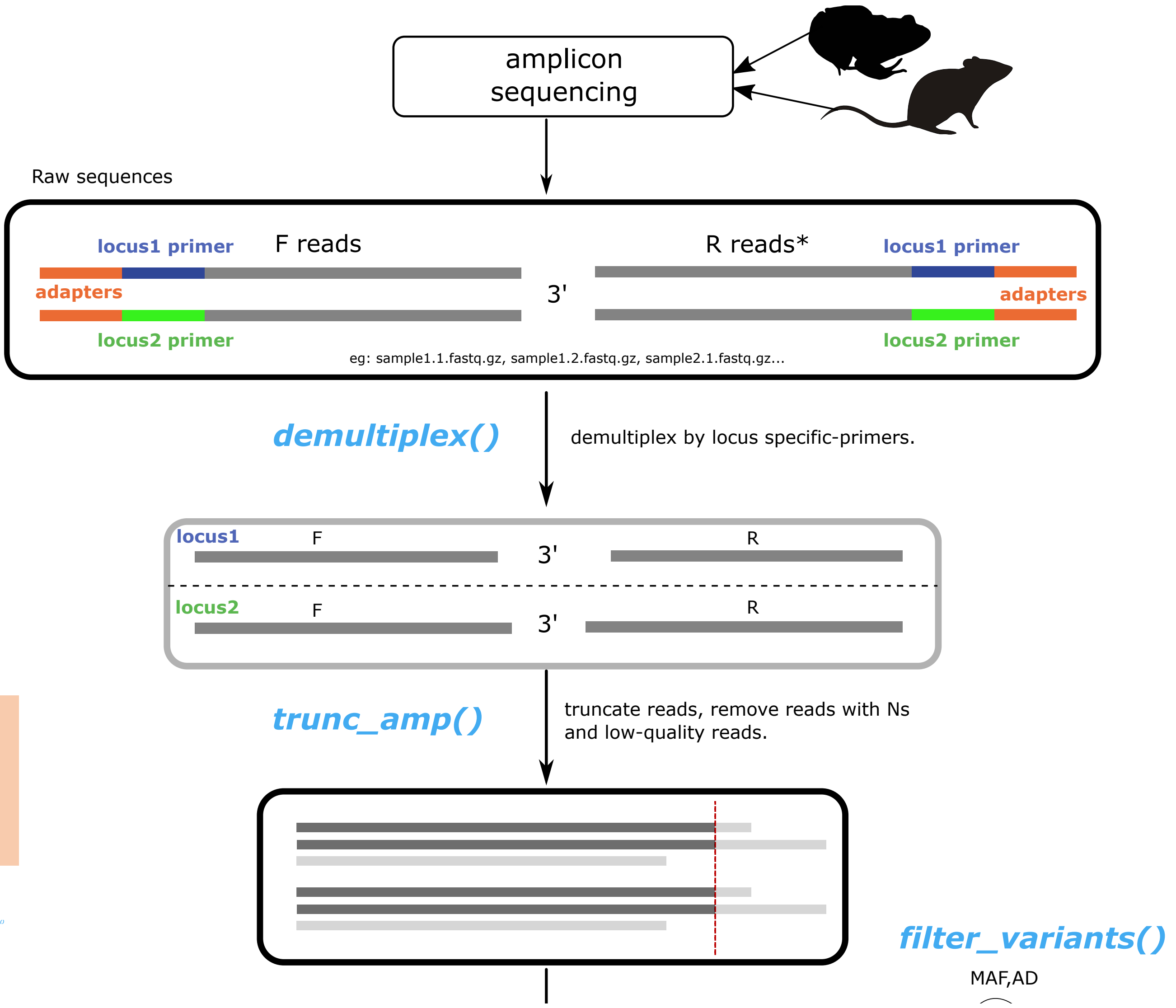
```
> var_seq2len(variants)
# A tibble: 1,295 × 4
  locus sample  variant reads
  <fct> <chr>     <int> <int>
1 abcg8 BOR1061     390  1028
2 abcg8 BOR1063     390   988
3 abcg8 BOR1069     390   955
4 abcg8 BOR1070     390   774
```

```
> genotypes
# A tibble: 2,372 × 8
  sample  locus  allele allele_no reads    nt md5                              sequence
  <chr>   <chr>  <chr>      <int> <dbl> <int> <chr>                            <chr>
1 BOR1061 abcg8  01             1  514    390 c3eab42bea56937040aa79a6bbea2724 TTGCCCACCCTGTTCATCCAT…
2 BOR1061 abcg8  01             2  514    390 c3eab42bea56937040aa79a6bbea2724 TTGCCCACCCTGTTCATCCAT…
3 BOR1061 alkbh7 01             1  196.   390 620057f2e547e688d1220db91a3a1d01 GCCTTCCTGGCCCCATCCCCT…
4 BOR1061 alkbh7 01             2  196.   390 620057f2e547e688d1220db91a3a1d01 GCCTTCCTGGCCCCATCCCCT…
5 BOR1061 apeh17 001            1 1374.   415 394a483d3ae2ce5605f1674b2400986c AAAGCCAGTGGAGCCACGATA…
```

```
>sampleA_locus1_a
ATCT...
>sampleA_locus1_a
ATCT...
...

>sampleB_locus1_a
ATCT...
>sampleB_locus1_b
ATGT...
...
```

```
> gen_tidy2wide(genotypes)
# A tibble: 44 × 28
  sample  abcg8 alkbh7 apeh17
  <chr>   <chr> <chr>  <chr>
1 BOR1061 01/01 01/01  001/001
2 BOR1063 01/01 01/01  001/001
3 BOR1069 01/02 01/01  001/001
4 BOR1070 01/02 01/01  001/001
```

```
> gen_tidy2compact(genotypes)
# A tibble: 1,186 × 3
  sample  locus  genotype
  <chr>   <chr>  <chr>
1 BOR1061 abcg8  01/01
2 BOR1061 alkbh7 01/01
3 BOR1061 apeh17 001/001
4 BOR1061 cd27   001/001
5 BOR1061 chrna9 01/02
```



amplicon sequencing

Raw sequences

locus1 primer

F reads

R reads*

locus1 primer

adapters

3'

adapters

locus2 primer

locus2 primer

eg: sample1.1.fastq.gz, sample1.2.fastq.gz, sample2.1.fastq.gz...

demultiplex()

demultiplex by locus specific-primers.

locus1

F

3'

R

locus2

F

3'

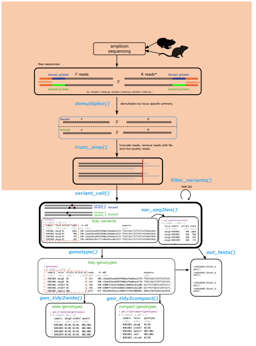
R

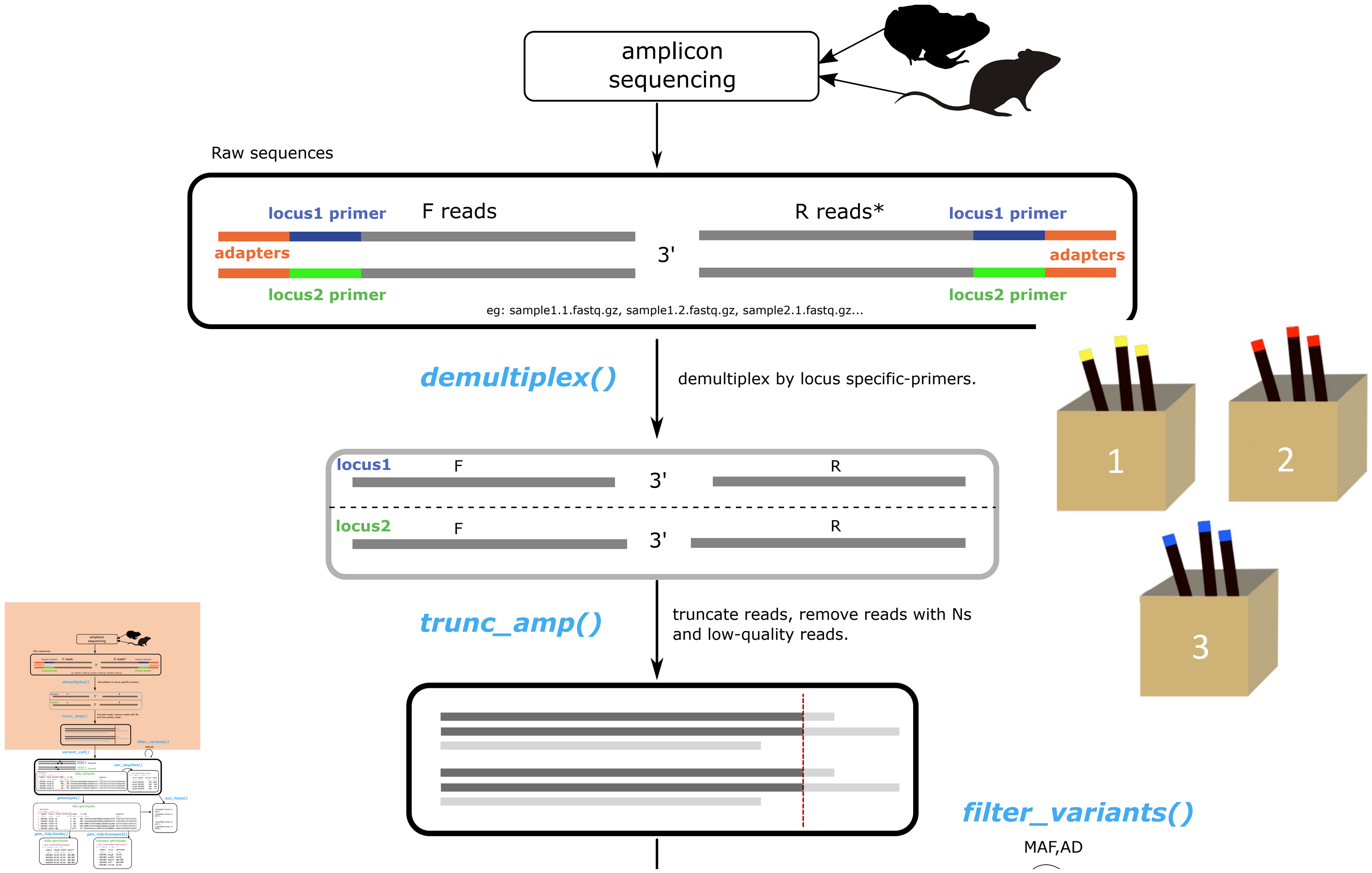
trunc_amp()

truncate reads, remove reads with Ns and low-quality reads.

filter_variants()

MAF,AD





amplicon sequencing

Raw sequences

locus1 primer

F reads

R reads*

locus1 primer

adapters

3'

adapters

locus2 primer

locus2 primer

eg: sample1.1.fastq.gz, sample1.2.fastq.gz, sample2.1.fastq.gz...

demultiplex()

demultiplex by locus specific-primers.

locus1

F

3'

R

locus2

F

3'

R

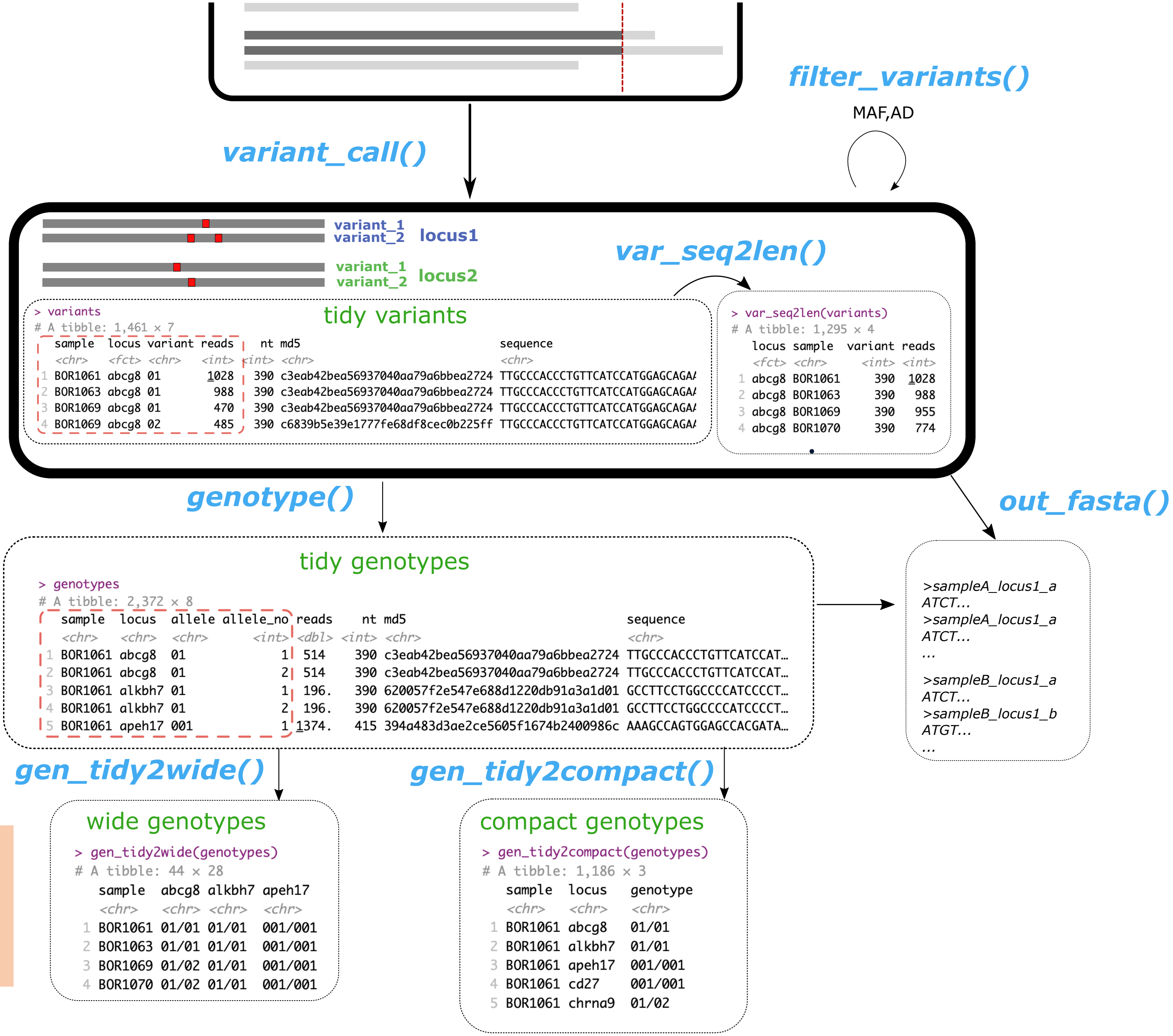
trunc_amp()

truncate reads, remove reads with Ns and low-quality reads.

filter_variants()

MAF,AD

RqueR



```
> variants
# A tibble: 1,461 × 7
  sample  locus variant reads    nt md5                              sequence
  <chr>   <fct> <chr>   <int> <int> <chr>                            <chr>
1 BOR1061 abcg8 01       1028   390 c3eab42bea56937040aa79a6bbea2724 TTGCCCACCCTGTTCATCCATGGAGCAGAA
2 BOR1063 abcg8 01        988   390 c3eab42bea56937040aa79a6bbea2724 TTGCCCACCCTGTTCATCCATGGAGCAGAA
3 BOR1069 abcg8 01        470   390 c3eab42bea56937040aa79a6bbea2724 TTGCCCACCCTGTTCATCCATGGAGCAGAA
4 BOR1069 abcg8 02        485   390 c6839b5e39e1777fe68df8cec0b225ff TTGCCCACCCTGTTCATCCATGGAGCAGAA
```

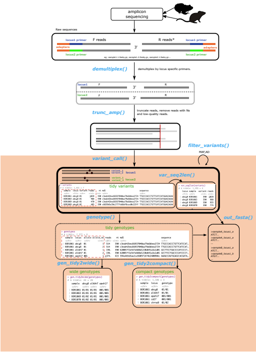
```
> var_seq2len(variants)
# A tibble: 1,295 × 4
  locus sample  variant reads
  <fct> <chr>     <int> <int>
1 abcg8 BOR1061     390  1028
2 abcg8 BOR1063     390   988
3 abcg8 BOR1069     390   955
4 abcg8 BOR1070     390   774
```

```
> genotypes
# A tibble: 2,372 × 8
  sample  locus  allele allele_no reads    nt md5                              sequence
  <chr>   <chr>  <chr>      <int> <dbl> <int> <chr>                            <chr>
1 BOR1061 abcg8  01             1  514    390 c3eab42bea56937040aa79a6bbea2724 TTGCCCACCCTGTTCATCCAT…
2 BOR1061 abcg8  01             2  514    390 c3eab42bea56937040aa79a6bbea2724 TTGCCCACCCTGTTCATCCAT…
3 BOR1061 alkbh7 01             1  196.   390 620057f2e547e688d1220db91a3a1d01 GCCTTCCTGGCCCCATCCCCT…
4 BOR1061 alkbh7 01             2  196.   390 620057f2e547e688d1220db91a3a1d01 GCCTTCCTGGCCCCATCCCCT…
5 BOR1061 apeh17 001            1 1374.   415 394a483d3ae2ce5605f1674b2400986c AAAGCCAGTGGAGCCACGATA…
```

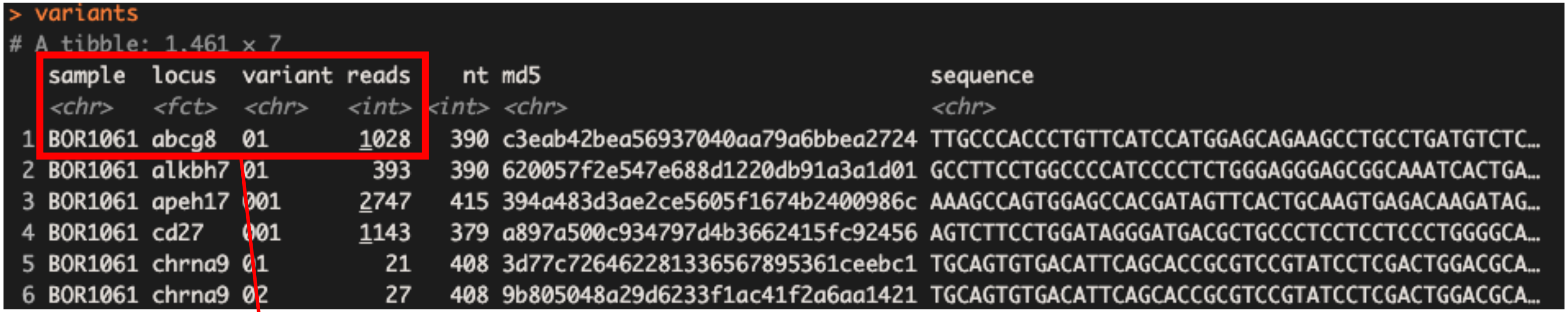
```
>sampleA_locus1_a
ATCT…
>sampleA_locus1_a
ATCT…
…

>sampleB_locus1_a
ATCT…
>sampleB_locus1_b
ATGT…
…
```

```
> gen_tidy2wide(genotypes)
# A tibble: 44 × 28
  sample  abcg8 alkbh7 apeh17
  <chr>   <chr> <chr>  <chr>
1 BOR1061 01/01 01/01  001/001
2 BOR1063 01/01 01/01  001/001
3 BOR1069 01/02 01/01  001/001
4 BOR1070 01/02 01/01  001/001
```

```
> gen_tidy2compact(genotypes)
# A tibble: 1,186 × 3
  sample  locus  genotype
  <chr>   <chr>  <chr>
1 BOR1061 abcg8  01/01
2 BOR1061 alkbh7 01/01
3 BOR1061 apeh17 001/001
4 BOR1061 cd27   001/001
5 BOR1061 chrna9 01/02
```

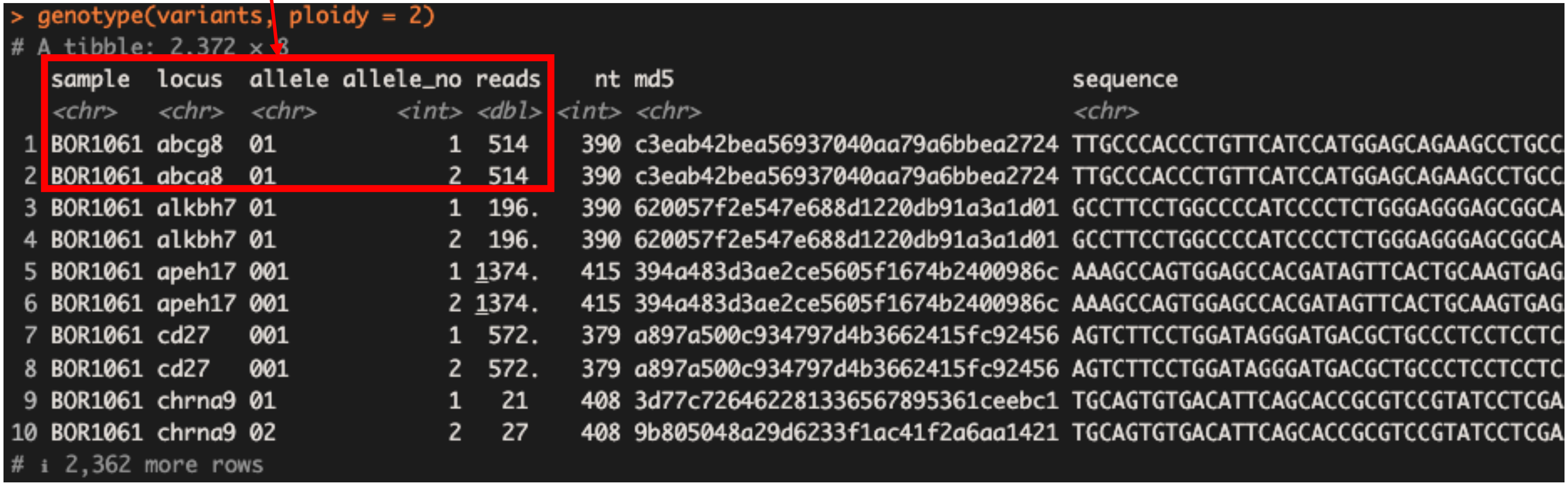


RqueR



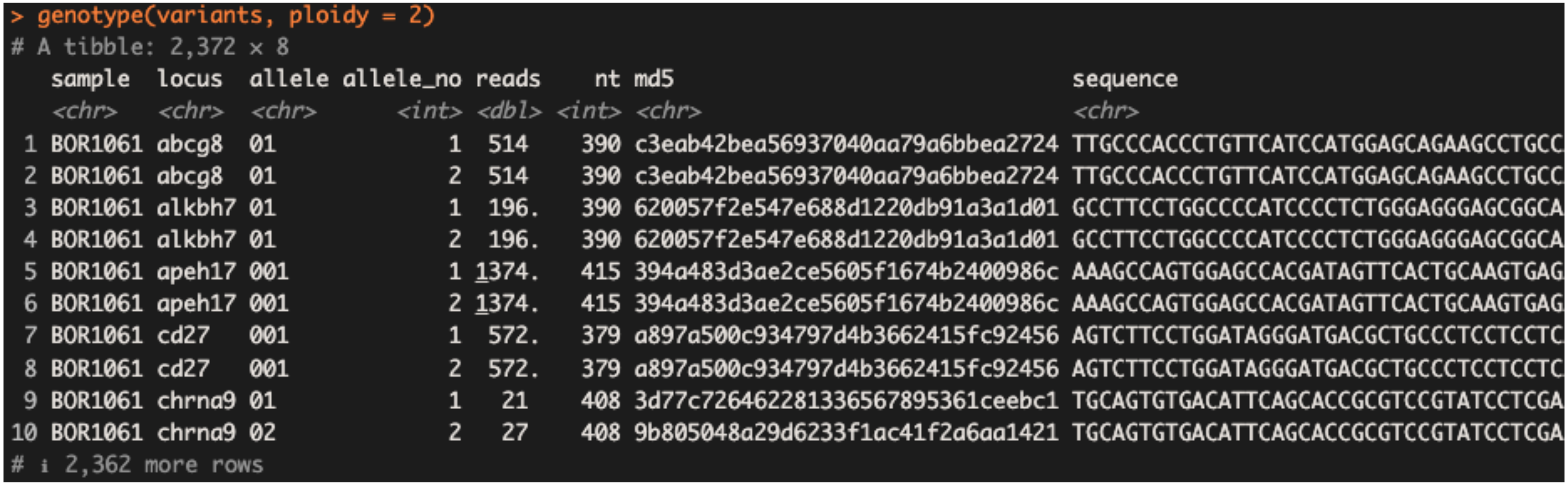
```
> variants
# A tibble: 1,461 × 7
  sample  locus  variant reads    nt md5                              sequence
  <chr>   <fct>  <chr>   <int> <int> <chr>                            <chr>
1 BOR1061 abcg8  01       1028   390 c3eab42bea56937040aa79a6bbea2724 TTGCCCACCCTGTTCATCCATGGAGCAGAAGCCTGCCTGATGTCTC…
2 BOR1061 alkbh7 01        393   390 620057f2e547e688d1220db91a3a1d01 GCCTTCCTGGCCCCATCCCCTCTGGGAGGGAGCGGCAAATCACTGA…
3 BOR1061 apeh17 001      2747   415 394a483d3ae2ce5605f1674b2400986c AAAGCCAGTGGAGCCACGATAGTTCACTGCAAGTGAGACAAGATAG…
4 BOR1061 cd27   001      1143   379 a897a500c934797d4b3662415fc92456 AGTCTTCCTGGATAGGGATGACGCTGCCCTCCTCCTCCCTGGGGCA…
5 BOR1061 chrna9 01          21   408 3d77c726462281336567895361ceebc1 TGCAGTGTGACATTCAGCACCGCGTCCGTATCCTCGACTGGACGCA…
6 BOR1061 chrna9 02          27   408 9b805048a29d6233f1ac41f2a6aa1421 TGCAGTGTGACATTCAGCACCGCGTCCGTATCCTCGACTGGACGCA…
```

genotype(ploidy = 2)



```
> genotype(variants, ploidy = 2)
# A tibble: 2,372 × 8
   sample  locus  allele allele_no reads    nt md5                              sequence
   <chr>   <chr>  <chr>      <int> <dbl> <int> <chr>                            <chr>
 1 BOR1061 abcg8  01             1  514    390 c3eab42bea56937040aa79a6bbea2724 TTGCCCACCCTGTTCATCCATGGAGCAGAAGCCTGCC
 2 BOR1061 abcg8  01             2  514    390 c3eab42bea56937040aa79a6bbea2724 TTGCCCACCCTGTTCATCCATGGAGCAGAAGCCTGCC
 3 BOR1061 alkbh7 01             1  196.   390 620057f2e547e688d1220db91a3a1d01 GCCTTCCTGGCCCCATCCCCTCTGGGAGGGAGCGGCA
 4 BOR1061 alkbh7 01             2  196.   390 620057f2e547e688d1220db91a3a1d01 GCCTTCCTGGCCCCATCCCCTCTGGGAGGGAGCGGCA
 5 BOR1061 apeh17 001            1 1374.   415 394a483d3ae2ce5605f1674b2400986c AAAGCCAGTGGAGCCACGATAGTTCACTGCAAGTGAG
 6 BOR1061 apeh17 001            2 1374.   415 394a483d3ae2ce5605f1674b2400986c AAAGCCAGTGGAGCCACGATAGTTCACTGCAAGTGAG
 7 BOR1061 cd27   001            1  572.   379 a897a500c934797d4b3662415fc92456 AGTCTTCCTGGATAGGGATGACGCTGCCCTCCTCCTC
 8 BOR1061 cd27   001            2  572.   379 a897a500c934797d4b3662415fc92456 AGTCTTCCTGGATAGGGATGACGCTGCCCTCCTCCTC
 9 BOR1061 chrna9 01             1   21    408 3d77c726462281336567895361ceebc1 TGCAGTGTGACATTCAGCACCGCGTCCGTATCCTCGA
10 BOR1061 chrna9 02             2   27    408 9b805048a29d6233f1ac41f2a6aa1421 TGCAGTGTGACATTCAGCACCGCGTCCGTATCCTCGA
# ℹ 2,362 more rows
```

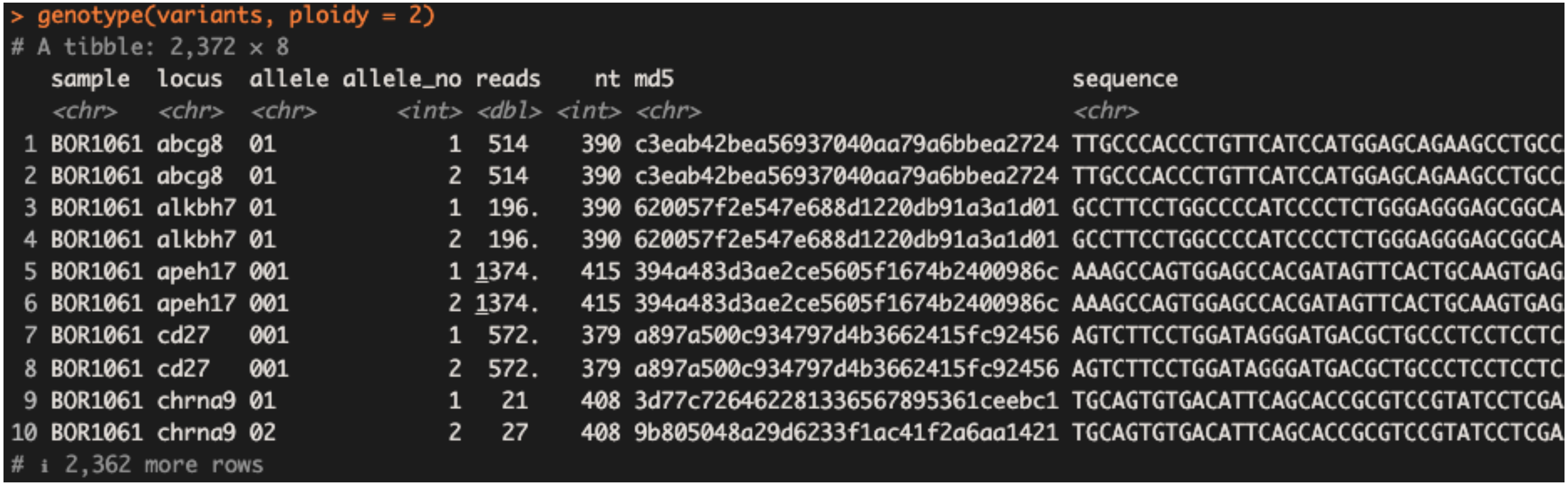
TIDY multilocus amplicon genotypes in R



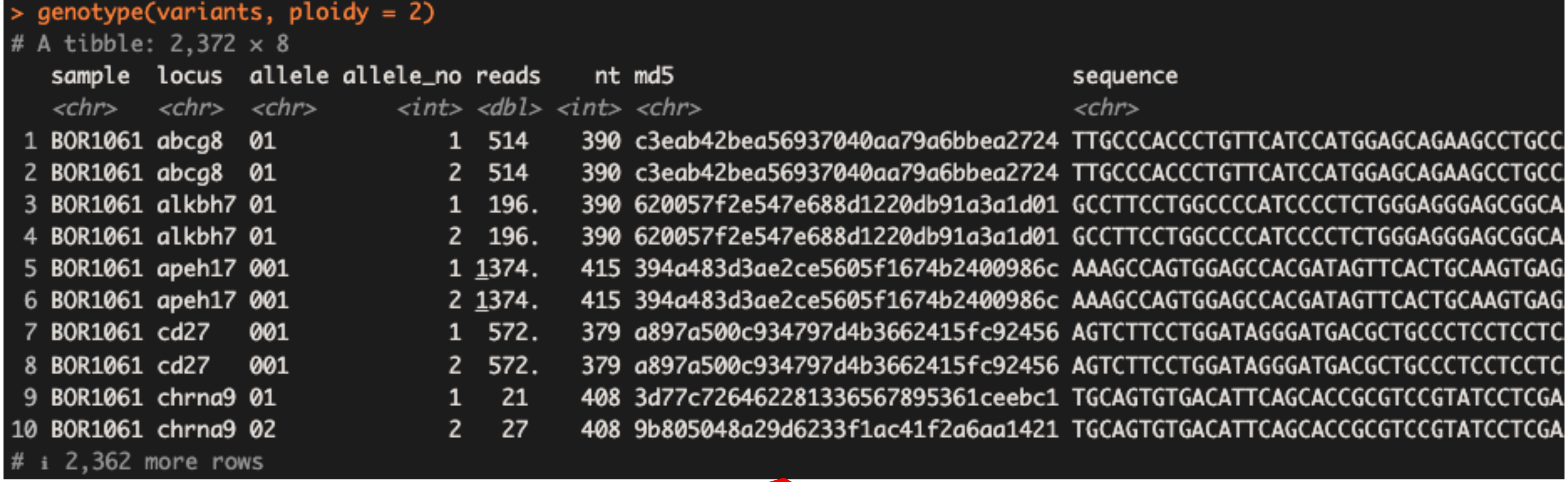
```
> genotype(variants, ploidy = 2)
# A tibble: 2,372 × 8
   sample  locus  allele allele_no reads    nt md5                              sequence
   <chr>   <chr>  <chr>      <int> <dbl> <int> <chr>                            <chr>
 1 BOR1061 abcg8  01             1  514    390 c3eab42bea56937040aa79a6bbea2724 TTGCCCACCCTGTTCATCCATGGAGCAGAAGCCTGCC
 2 BOR1061 abcg8  01             2  514    390 c3eab42bea56937040aa79a6bbea2724 TTGCCCACCCTGTTCATCCATGGAGCAGAAGCCTGCC
 3 BOR1061 alkbh7 01             1  196.   390 620057f2e547e688d1220db91a3a1d01 GCCTTCCTGGCCCCATCCCCTCTGGGAGGGAGCGGCA
 4 BOR1061 alkbh7 01             2  196.   390 620057f2e547e688d1220db91a3a1d01 GCCTTCCTGGCCCCATCCCCTCTGGGAGGGAGCGGCA
 5 BOR1061 apeh17 001            1 1374.   415 394a483d3ae2ce5605f1674b2400986c AAAGCCAGTGGAGCCACGATAGTTCACTGCAAGTGAG
 6 BOR1061 apeh17 001            2 1374.   415 394a483d3ae2ce5605f1674b2400986c AAAGCCAGTGGAGCCACGATAGTTCACTGCAAGTGAG
 7 BOR1061 cd27   001            1  572.   379 a897a500c934797d4b3662415fc92456 AGTCTTCCTGGATAGGGATGACGCTGCCCTCCTCCTC
 8 BOR1061 cd27   001            2  572.   379 a897a500c934797d4b3662415fc92456 AGTCTTCCTGGATAGGGATGACGCTGCCCTCCTCCTC
 9 BOR1061 chrna9 01             1    21   408 3d77c726462281336567895361ceebc1 TGCAGTGTGACATTCAGCACCGCGTCCGTATCCTCGA
10 BOR1061 chrna9 02             2    27   408 9b805048a29d6233f1ac41f2a6aa1421 TGCAGTGTGACATTCAGCACCGCGTCCGTATCCTCGA
# ℹ 2,362 more rows
```

Ventajas formato ‘tidy’:

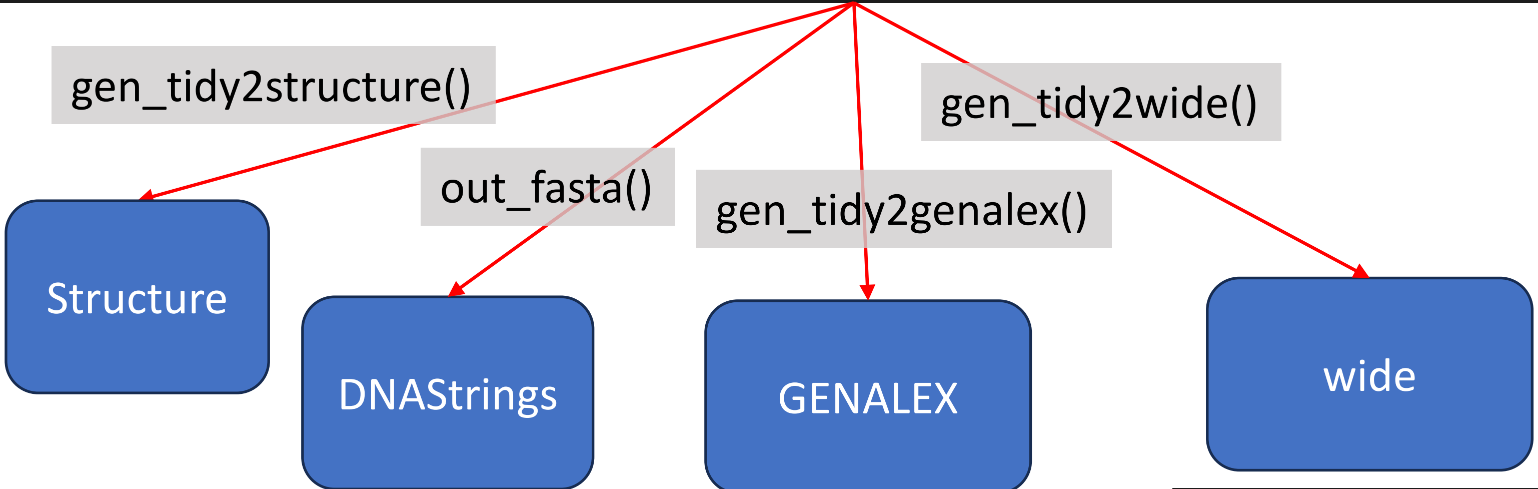
- escalado “row-wise” y “column-wise”
- manipulación (eg filtrado, selección)
- re-formateo

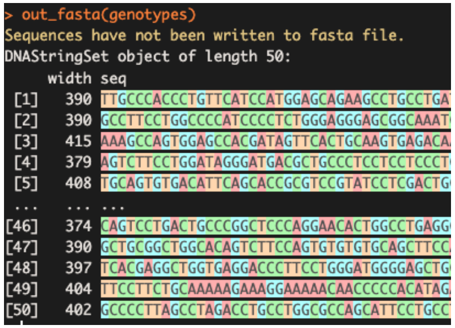


```
> genotype(variants, ploidy = 2)
# A tibble: 2,372 × 8
   sample  locus  allele allele_no reads    nt md5                              sequence
   <chr>   <chr>  <chr>      <int> <dbl> <int> <chr>                            <chr>
 1 BOR1061 abcg8  01             1  514    390 c3eab42bea56937040aa79a6bbea2724 TTGCCCACCCTGTTCATCCATGGAGCAGAAGCCTGCC
 2 BOR1061 abcg8  01             2  514    390 c3eab42bea56937040aa79a6bbea2724 TTGCCCACCCTGTTCATCCATGGAGCAGAAGCCTGCC
 3 BOR1061 alkbh7 01             1  196.   390 620057f2e547e688d1220db91a3a1d01 GCCTTCCTGGCCCCATCCCCTCTGGGAGGGAGCGGCA
 4 BOR1061 alkbh7 01             2  196.   390 620057f2e547e688d1220db91a3a1d01 GCCTTCCTGGCCCCATCCCCTCTGGGAGGGAGCGGCA
 5 BOR1061 apeh17 001            1 1374.   415 394a483d3ae2ce5605f1674b2400986c AAAGCCAGTGGAGCCACGATAGTTCACTGCAAGTGAG
 6 BOR1061 apeh17 001            2 1374.   415 394a483d3ae2ce5605f1674b2400986c AAAGCCAGTGGAGCCACGATAGTTCACTGCAAGTGAG
 7 BOR1061 cd27   001            1  572.   379 a897a500c934797d4b3662415fc92456 AGTCTTCCTGGATAGGGATGACGCTGCCCTCCTCCTC
 8 BOR1061 cd27   001            2  572.   379 a897a500c934797d4b3662415fc92456 AGTCTTCCTGGATAGGGATGACGCTGCCCTCCTCCTC
 9 BOR1061 chrna9 01             1   21    408 3d77c7264622813365678953б1ceebc1 TGCAGTGTGACATTCAGCACCGCGTCCGTATCCTCGA
10 BOR1061 chrna9 02             2   27    408 9b805048a29d6233f1ac41f2a6aa1421 TGCAGTGTGACATTCAGCACCGCGTCCGTATCCTCGA
# ℹ 2,362 more rows
```

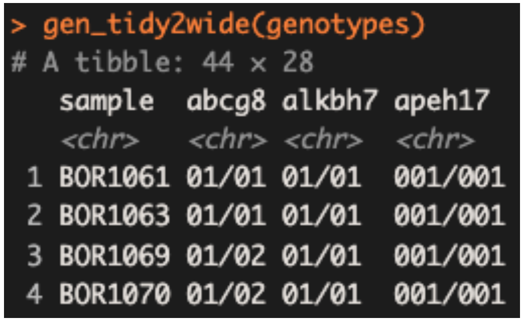


```
> genotype(variants, ploidy = 2)
# A tibble: 2,372 × 8
   sample  locus  allele allele_no reads    nt md5                              sequence
   <chr>   <chr>  <chr>      <int> <dbl> <int> <chr>                            <chr>
 1 BOR1061 abcg8  01             1  514    390 c3eab42bea56937040aa79a6bbea2724 TTGCCCACCCTGTTCATCCATGGAGCAGAAGCCTGCC
 2 BOR1061 abcg8  01             2  514    390 c3eab42bea56937040aa79a6bbea2724 TTGCCCACCCTGTTCATCCATGGAGCAGAAGCCTGCC
 3 BOR1061 alkbh7 01             1  196.   390 620057f2e547e688d1220db91a3a1d01 GCCTTCCTGGCCCCATCCCCTCTGGGAGGGAGCGGCA
 4 BOR1061 alkbh7 01             2  196.   390 620057f2e547e688d1220db91a3a1d01 GCCTTCCTGGCCCCATCCCCTCTGGGAGGGAGCGGCA
 5 BOR1061 apeh17 001            1 1374.   415 394a483d3ae2ce5605f1674b2400986c AAAGCCAGTGGAGCCACGATAGTTCACTGCAAGTGAG
 6 BOR1061 apeh17 001            2 1374.   415 394a483d3ae2ce5605f1674b2400986c AAAGCCAGTGGAGCCACGATAGTTCACTGCAAGTGAG
 7 BOR1061 cd27   001            1  572.   379 a897a500c934797d4b3662415fc92456 AGTCTTCCTGGATAGGGATGACGCTGCCCTCCTCCTC
 8 BOR1061 cd27   001            2  572.   379 a897a500c934797d4b3662415fc92456 AGTCTTCCTGGATAGGGATGACGCTGCCCTCCTCCTC
 9 BOR1061 chrna9 01             1   21    408 3d77c7264622813365678953б1ceebc1 TGCAGTGTGACATTCAGCACCGCGTCCGTATCCTCGA
10 BOR1061 chrna9 02             2   27    408 9b805048a29d6233f1ac41f2a6aa1421 TGCAGTGTGACATTCAGCACCGCGTCCGTATCCTCGA
# ℹ 2,362 more rows
```





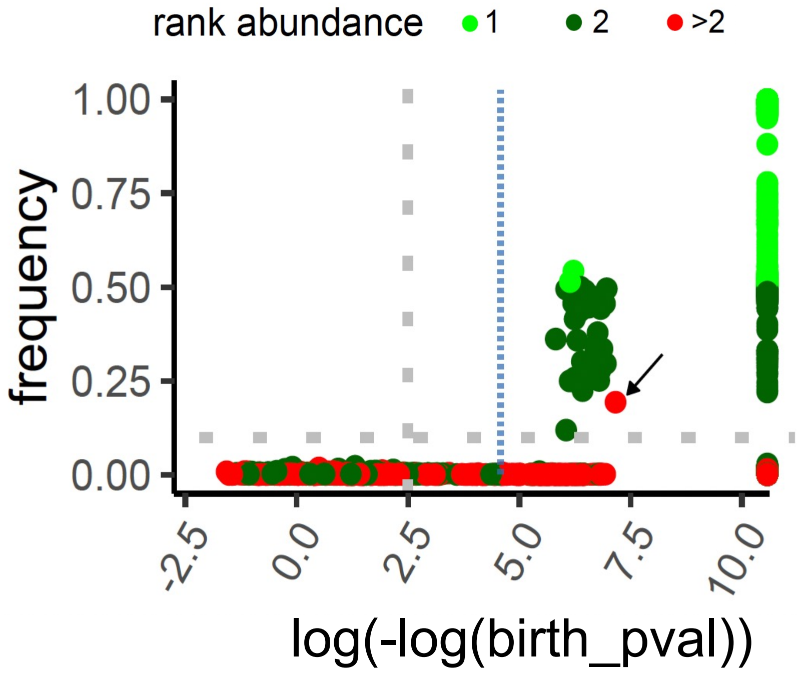
```
> out_fasta(genotypes)
Sequences have not been written to fasta file.
DNAStringSet object of length 50:
     width seq
 [1]   390 TTGCCCACCCTGTTCATCCATGGAGCAGAAGCCTGCCTGA
 [2]   390 GCCTTCCTGGCCCCATCCCCTCTGGGAGGGAGCGGCAAAT
 [3]   415 AAAGCCAGTGGAGCCACGATAGTTCACTGCAAGTGAGACA
 [4]   379 AGTCTTCCTGGATAGGGATGACGCTGCCCTCCTCCTCCCT
 [5]   408 TGCAGTGTGACATTCAGCACCGCGTCCGTATCCTCGACTG
 ...   ... ...
[46]   374 CAGTCCTGACTGCCCGGCTCCCAGGAACACTGGCCTGAGG
[47]   390 GCTGCGGCTGGCACAGTCTTCCAGTGTGTGTGCAGCTTCC
[48]   397 TCACGAGGCTGGTGAGGACCCTTCCTGGGATGGGGAGCTG
[49]   404 TTCCTTCTGCAAAAAGAAAGGAAAAACAACCCCCACATAG
[50]   402 GCCCCTTAGCCTAGACCTGCCTGGCGCCAGCATTCCTGCC
```



```
> gen_tidy2wide(genotypes)
# A tibble: 44 × 28
  sample  abcg8 alkbh7 apeh17
  <chr>   <chr> <chr>  <chr>
1 BOR1061 01/01 01/01  001/001
2 BOR1063 01/01 01/01  001/001
3 BOR1069 01/02 01/01  001/001
4 BOR1070 01/02 01/01  001/001
```

Diagnosis

explore_dada()

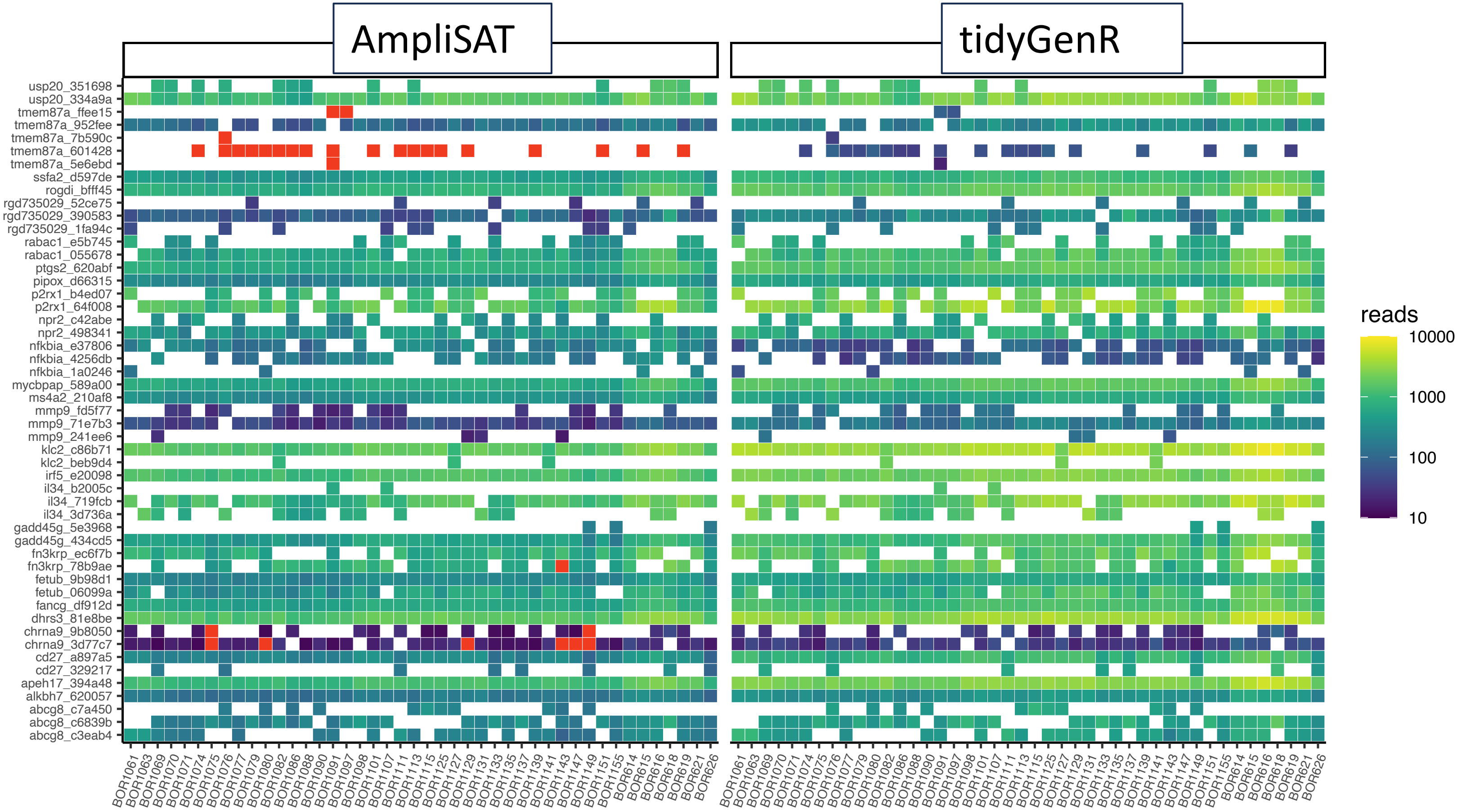


probabilidad de variante real

IIIRqueR

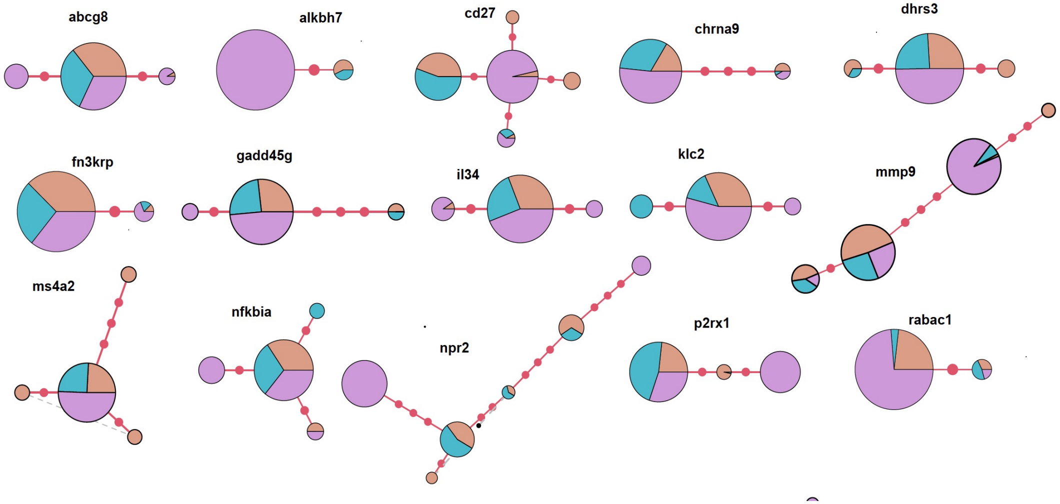
Diagnosis

compare_calls()



Caso práctico

- Amplicones 27 intrones
- Illumina MiSeq 300PE.
- 44 muestras de *Rattus baluensis*, de Borneo.



Adaptado de Greenwood et al. (2011) (https://doi.org/10.1371/journal.pone.0021114) (CC BY 4.0)

Conclusiones

- alternativa fiable para determinar variantes de librerías multilocus de amplicones.
- flexible, admite diferentes puntos de entrada, y ajustar parámetros.
- datos ordenados 'tidy' fáciles de manejar.
- funciones de diagnóstico.

Estado de desarrollo

- código | | | | | | | | | | 100 %
- documentación | | | | | | | | | | 100 %
- chequeos | | | | | | | | | | 80 %
- manuscrito | | | | | | | | | | 70 %

https://github.com/csmiguel/tidyGenR/ (PRIVADO)

Agradecimientos

- Ministerio de Economía y Competitividad CGL2014-58793-P y PID2020-120115GB-100
- Anna Cornellas
- Arlo Hinckley
- “Laboratorio de Ecología Molecular” (LEM-EBD), Doñana ICTS-RBD

Jennifer Leonard

GOBIERNO DE ESPAÑA | MINISTERIO DE CIENCIA E INNOVACIÓN

CSIC CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

Estación Biológica Doñana CSIC

Estación Biológica Doñana 60 ANIVERSARIO

RqueR