

Tarea 8.3

César Emiliano Escalona Prado

Variables aplicadas

Distancia permitida entre stacks (M) se empleó su valor por default (2).

Profundidad mínima del stack/ profundidad mínima de cobertura (m) se empleó su valor por default (3).

Distancia permitida entre loci de catálogo (n), se emplearon los valores 1, 2 y 3. La variable n actúa en el desarrollo del programa cstacks, éste emplea los locis generados a partir del ustacks, los cuales fusiona en un catálogo que debe contener todos los loci y alelos de la población. En el caso de que este parámetro se establezca con un valor demasiado bajo, habrá loci entre los individuos que están representados independientemente en el catálogo que son realmente el mismo locus. Si establece este parámetro demasiado alto, permitirá que los loci se unan en el espacio de secuencia para encadenarse y crear loci grandes y erróneos en el catálogo.

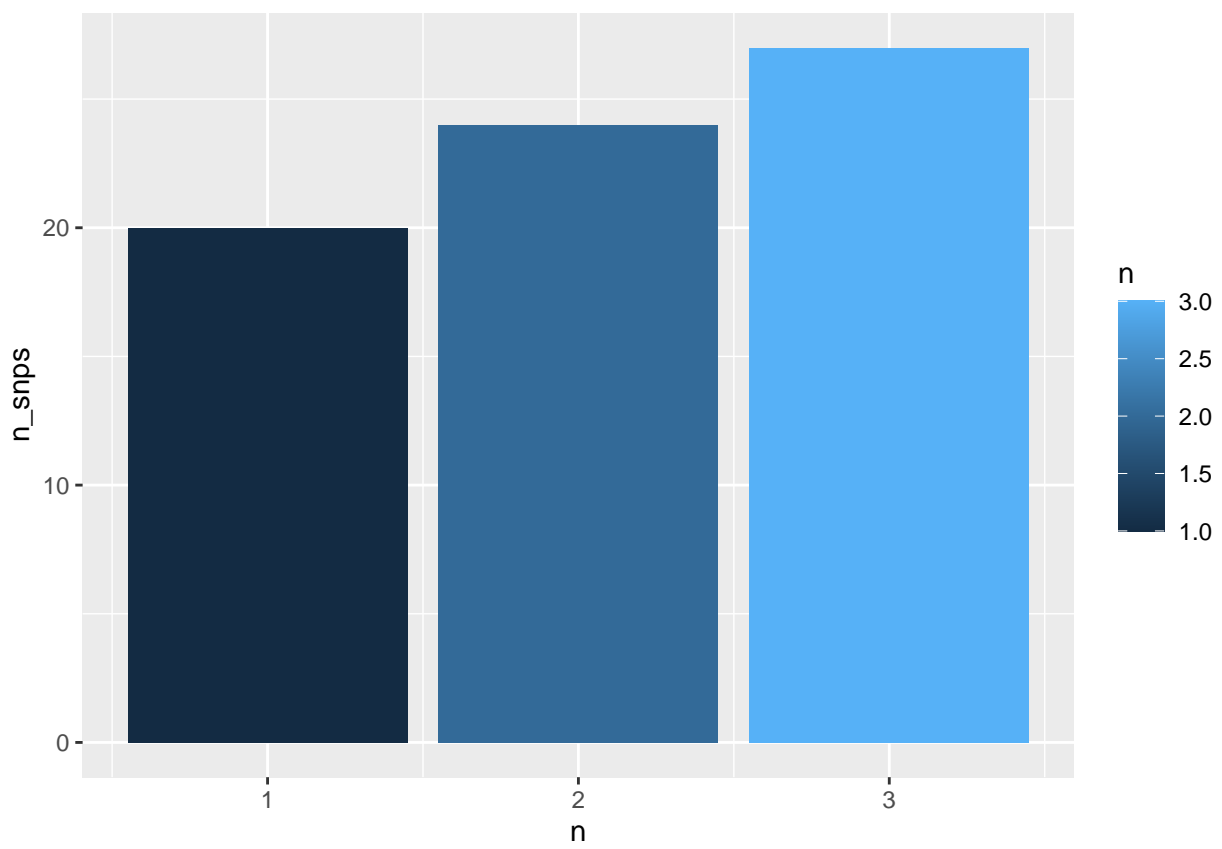
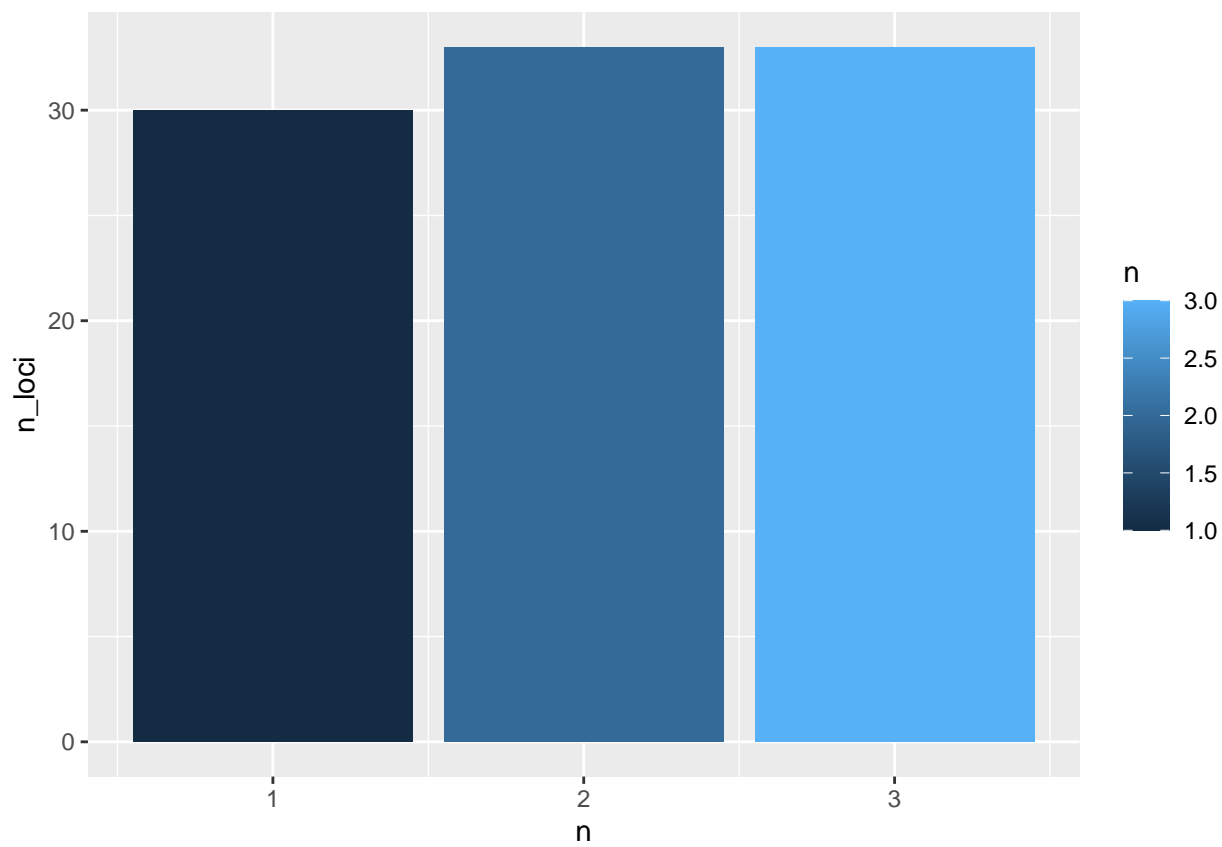
Resultados

n1) Se eliminaron 50508 de 50538 loci que no pasaron las restricciones de muestra / población. Se mantuvieron 30 loci, compuestos por 2852 sitios; 2 de esos sitios fueron filtrados, quedaron 20 sitios variantes. Sitios genotipados en promedio por locus: 95.33bp (stderr 0.20). Sitios genotipados en promedio por locus: 95.07bp (stderr 0.07)

n2) Se eliminaron 49225 de 49258 loci que no pasaron las restricciones de muestra / población. Se mantuvieron 33 loci, compuestos por 3146 sitios; Se filtraron 2 de esos sitios, quedaron 24 sitios variantes. Sitios genotipados en promedio por locus: 95.33bp (stderr 0.20).

n3) Se eliminaron 48511 de 48544 loci que no pasaron las restricciones de muestra / población. Se mantuvieron 33 loci, compuestos por 3146 sitios; 2 de esos sitios fueron filtrados, quedaron 27 sitios variantes. Sitios genotipados en promedio por locus: 95.33bp (stderr 0.20).

M	n	m	n_loci	n_snps
2	1	3	30	20
2	2	3	33	24
2	3	3	33	27



```

d= read.csv('./n_snps_per_locus.tsv')

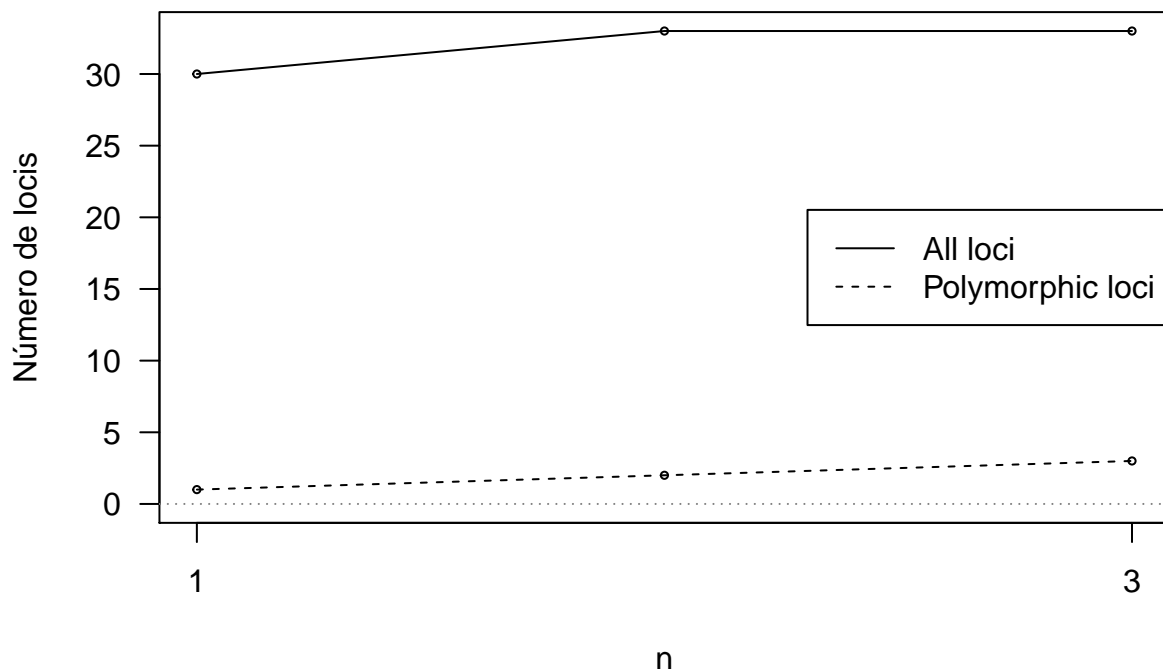
plot(NULL,
      xlim=range(d$n),
      ylim=range(c(0, d$n_loci)),
      xlab='n',
      ylab='Número de locis',
      main='Número de locis muestreados (80%) conforme n aumenta',
      xaxt='n',
      las=2
    )
abline(h=0:20*5000, lty='dotted', col='grey50')
axis(1, at=c(1,3,5,7,9))
legend('right', c('All loci', 'Polymorphic loci'), lty=c('solid', 'dashed'))

# Total number of loci.
lines(d$n, d$n_loci)
points(d$n, d$n_loci, cex=0.5)

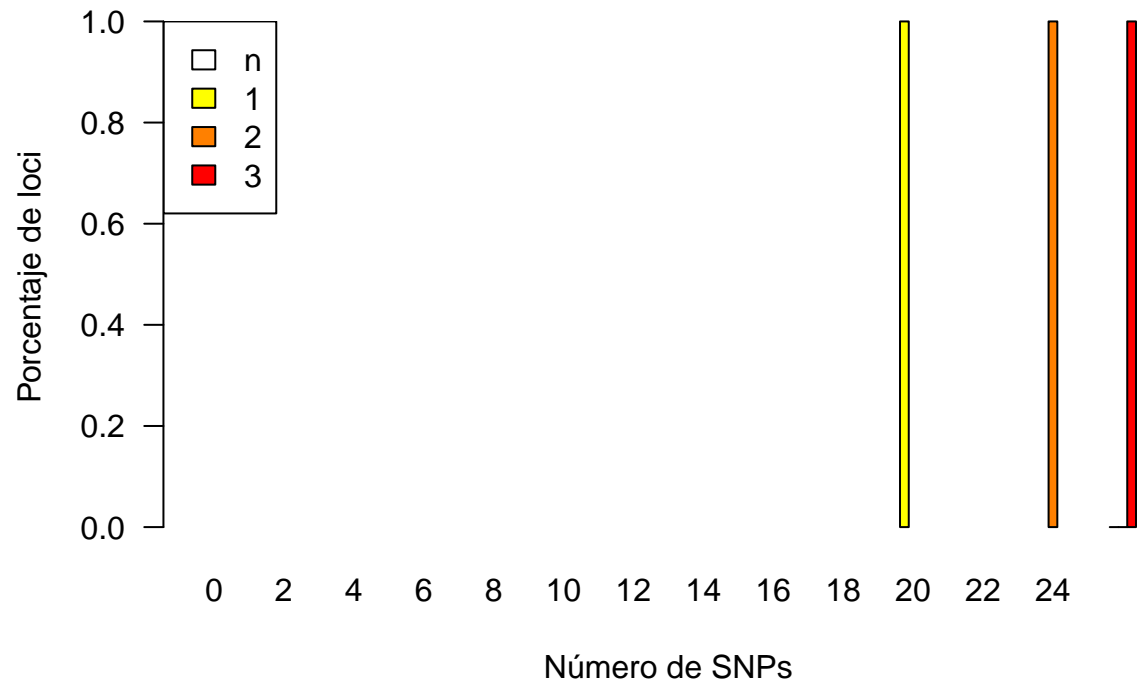
# Number of polymorphic loci.
lines(d$n, d$n_loci_poly, lty='dashed')
points(d$n, d$n_loci_poly, cex=0.5)

```

Número de locis muestreados (80%) conforme n aumenta



Distribución del número de SNPs por locus para un rango de valores de n



Usar $n=3$ da en mayor número de sitios variantes, manteniendo el mayor número de loci, por lo cual sería el valor que emplearía de n .