Filtering VCF files

Hazel Daniels (adapted from Javier Tabima)

30 March 2021

# Reading in data

vcfNA <- read.vcfR("ramorumNA1.vcf.gz")

## Scanning file to determine attributes.  
## File attributes:  
## meta lines: 66  
## header\_line: 67  
## variant count: 484461  
## column count: 144  
## Meta line 66 read in.  
## All meta lines processed.  
## gt matrix initialized.  
## Character matrix gt created.  
## Character matrix gt rows: 484461  
## Character matrix gt cols: 144  
## skip: 0  
## nrows: 484461  
## row\_num: 0  
## Processed variant 1000Processed variant 2000Processed variant 3000Processed variant 4000Processed variant 5000Processed variant 6000Processed variant 7000Processed variant 8000Processed variant 9000Processed variant 10000Processed variant 11000Processed variant 12000Processed variant 13000Processed variant 14000Processed variant 15000Processed variant 16000Processed variant 17000Processed variant 18000Processed variant 19000Processed variant 20000Processed variant 21000Processed variant 22000Processed variant 23000Processed variant 24000Processed variant 25000Processed variant 26000Processed variant 27000Processed variant 28000Processed variant 29000Processed variant 30000Processed variant 31000Processed variant 32000Processed variant 33000Processed variant 34000Processed variant 35000Processed variant 36000Processed variant 37000Processed variant 38000Processed variant 39000Processed variant 40000Processed variant 41000Processed variant 42000Processed variant 43000Processed variant 44000Processed variant 45000Processed variant 46000Processed variant 47000Processed variant 48000Processed variant 49000Processed variant 50000Processed variant 51000Processed variant 52000Processed variant 53000Processed variant 54000Processed variant 55000Processed variant 56000Processed variant 57000Processed variant 58000Processed variant 59000Processed variant 60000Processed variant 61000Processed variant 62000Processed variant 63000Processed variant 64000Processed variant 65000Processed variant 66000Processed variant 67000Processed variant 68000Processed variant 69000Processed variant 70000Processed variant 71000Processed variant 72000Processed variant 73000Processed variant 74000Processed variant 75000Processed variant 76000Processed variant 77000Processed variant 78000Processed variant 79000Processed variant 80000Processed variant 81000Processed variant 82000Processed variant 83000Processed variant 84000Processed variant 85000Processed variant 86000Processed variant 87000Processed variant 88000Processed variant 89000Processed variant 90000Processed variant 91000Processed variant 92000Processed variant 93000Processed variant 94000Processed variant 95000Processed variant 96000Processed variant 97000Processed variant 98000Processed variant 99000Processed variant 100000Processed variant 101000Processed variant 102000Processed variant 103000Processed variant 104000Processed variant 105000Processed variant 106000Processed variant 107000Processed variant 108000Processed variant 109000Processed variant 110000Processed variant 111000Processed variant 112000Processed variant 113000Processed variant 114000Processed variant 115000Processed variant 116000Processed variant 117000Processed variant 118000Processed variant 119000Processed variant 120000Processed variant 121000Processed variant 122000Processed variant 123000Processed variant 124000Processed variant 125000Processed variant 126000Processed variant 127000Processed variant 128000Processed variant 129000Processed variant 130000Processed variant 131000Processed variant 132000Processed variant 133000Processed variant 134000Processed variant 135000Processed variant 136000Processed variant 137000Processed variant 138000Processed variant 139000Processed variant 140000Processed variant 141000Processed variant 142000Processed variant 143000Processed variant 144000Processed variant 145000Processed variant 146000Processed variant 147000Processed variant 148000Processed variant 149000Processed variant 150000Processed variant 151000Processed variant 152000Processed variant 153000Processed variant 154000Processed variant 155000Processed variant 156000Processed variant 157000Processed variant 158000Processed variant 159000Processed variant 160000Processed variant 161000Processed variant 162000Processed variant 163000Processed variant 164000Processed variant 165000Processed variant 166000Processed variant 167000Processed variant 168000Processed variant 169000Processed variant 170000Processed variant 171000Processed variant 172000Processed variant 173000Processed variant 174000Processed variant 175000Processed variant 176000Processed variant 177000Processed variant 178000Processed variant 179000Processed variant 180000Processed variant 181000Processed variant 182000Processed variant 183000Processed variant 184000Processed variant 185000Processed variant 186000Processed variant 187000Processed variant 188000Processed variant 189000Processed variant 190000Processed variant 191000Processed variant 192000Processed variant 193000Processed variant 194000Processed variant 195000Processed variant 196000Processed variant 197000Processed variant 198000Processed variant 199000Processed variant 200000Processed variant 201000Processed variant 202000Processed variant 203000Processed variant 204000Processed variant 205000Processed variant 206000Processed variant 207000Processed variant 208000Processed variant 209000Processed variant 210000Processed variant 211000Processed variant 212000Processed variant 213000Processed variant 214000Processed variant 215000Processed variant 216000Processed variant 217000Processed variant 218000Processed variant 219000Processed variant 220000Processed variant 221000Processed variant 222000Processed variant 223000Processed variant 224000Processed variant 225000Processed variant 226000Processed variant 227000Processed variant 228000Processed variant 229000Processed variant 230000Processed variant 231000Processed variant 232000Processed variant 233000Processed variant 234000Processed variant 235000Processed variant 236000Processed variant 237000Processed variant 238000Processed variant 239000Processed variant 240000Processed variant 241000Processed variant 242000Processed variant 243000Processed variant 244000Processed variant 245000Processed variant 246000Processed variant 247000Processed variant 248000Processed variant 249000Processed variant 250000Processed variant 251000Processed variant 252000Processed variant 253000Processed variant 254000Processed variant 255000Processed variant 256000Processed variant 257000Processed variant 258000Processed variant 259000Processed variant 260000Processed variant 261000Processed variant 262000Processed variant 263000Processed variant 264000Processed variant 265000Processed variant 266000Processed variant 267000Processed variant 268000Processed variant 269000Processed variant 270000Processed variant 271000Processed variant 272000Processed variant 273000Processed variant 274000Processed variant 275000Processed variant 276000Processed variant 277000Processed variant 278000Processed variant 279000Processed variant 280000Processed variant 281000Processed variant 282000Processed variant 283000Processed variant 284000Processed variant 285000Processed variant 286000Processed variant 287000Processed variant 288000Processed variant 289000Processed variant 290000Processed variant 291000Processed variant 292000Processed variant 293000Processed variant 294000Processed variant 295000Processed variant 296000Processed variant 297000Processed variant 298000Processed variant 299000Processed variant 300000Processed variant 301000Processed variant 302000Processed variant 303000Processed variant 304000Processed variant 305000Processed variant 306000Processed variant 307000Processed variant 308000Processed variant 309000Processed variant 310000Processed variant 311000Processed variant 312000Processed variant 313000Processed variant 314000Processed variant 315000Processed variant 316000Processed variant 317000Processed variant 318000Processed variant 319000Processed variant 320000Processed variant 321000Processed variant 322000Processed variant 323000Processed variant 324000Processed variant 325000Processed variant 326000Processed variant 327000Processed variant 328000Processed variant 329000Processed variant 330000Processed variant 331000Processed variant 332000Processed variant 333000Processed variant 334000Processed variant 335000Processed variant 336000Processed variant 337000Processed variant 338000Processed variant 339000Processed variant 340000Processed variant 341000Processed variant 342000Processed variant 343000Processed variant 344000Processed variant 345000Processed variant 346000Processed variant 347000Processed variant 348000Processed variant 349000Processed variant 350000Processed variant 351000Processed variant 352000Processed variant 353000Processed variant 354000Processed variant 355000Processed variant 356000Processed variant 357000Processed variant 358000Processed variant 359000Processed variant 360000Processed variant 361000Processed variant 362000Processed variant 363000Processed variant 364000Processed variant 365000Processed variant 366000Processed variant 367000Processed variant 368000Processed variant 369000Processed variant 370000Processed variant 371000Processed variant 372000Processed variant 373000Processed variant 374000Processed variant 375000Processed variant 376000Processed variant 377000Processed variant 378000Processed variant 379000Processed variant 380000Processed variant 381000Processed variant 382000Processed variant 383000Processed variant 384000Processed variant 385000Processed variant 386000Processed variant 387000Processed variant 388000Processed variant 389000Processed variant 390000Processed variant 391000Processed variant 392000Processed variant 393000Processed variant 394000Processed variant 395000Processed variant 396000Processed variant 397000Processed variant 398000Processed variant 399000Processed variant 400000Processed variant 401000Processed variant 402000Processed variant 403000Processed variant 404000Processed variant 405000Processed variant 406000Processed variant 407000Processed variant 408000Processed variant 409000Processed variant 410000Processed variant 411000Processed variant 412000Processed variant 413000Processed variant 414000Processed variant 415000Processed variant 416000Processed variant 417000Processed variant 418000Processed variant 419000Processed variant 420000Processed variant 421000Processed variant 422000Processed variant 423000Processed variant 424000Processed variant 425000Processed variant 426000Processed variant 427000Processed variant 428000Processed variant 429000Processed variant 430000Processed variant 431000Processed variant 432000Processed variant 433000Processed variant 434000Processed variant 435000Processed variant 436000Processed variant 437000Processed variant 438000Processed variant 439000Processed variant 440000Processed variant 441000Processed variant 442000Processed variant 443000Processed variant 444000Processed variant 445000Processed variant 446000Processed variant 447000Processed variant 448000Processed variant 449000Processed variant 450000Processed variant 451000Processed variant 452000Processed variant 453000Processed variant 454000Processed variant 455000Processed variant 456000Processed variant 457000Processed variant 458000Processed variant 459000Processed variant 460000Processed variant 461000Processed variant 462000Processed variant 463000Processed variant 464000Processed variant 465000Processed variant 466000Processed variant 467000Processed variant 468000Processed variant 469000Processed variant 470000Processed variant 471000Processed variant 472000Processed variant 473000Processed variant 474000Processed variant 475000Processed variant 476000Processed variant 477000Processed variant 478000Processed variant 479000Processed variant 480000Processed variant 481000Processed variant 482000Processed variant 483000Processed variant 484000Processed variant: 484461  
## All variants processed

raw.vcfNA <- nrow(vcfNA)  
  
vcfEU <- read.vcfR("ramorumEU1.vcf.gz")

## Scanning file to determine attributes.  
## File attributes:  
## meta lines: 65  
## header\_line: 66  
## variant count: 499848  
## column count: 169  
## Meta line 65 read in.  
## All meta lines processed.  
## gt matrix initialized.  
## Character matrix gt created.  
## Character matrix gt rows: 499848  
## Character matrix gt cols: 169  
## skip: 0  
## nrows: 499848  
## row\_num: 0  
## Processed variant 1000Processed variant 2000Processed variant 3000Processed variant 4000Processed variant 5000Processed variant 6000Processed variant 7000Processed variant 8000Processed variant 9000Processed variant 10000Processed variant 11000Processed variant 12000Processed variant 13000Processed variant 14000Processed variant 15000Processed variant 16000Processed variant 17000Processed variant 18000Processed variant 19000Processed variant 20000Processed variant 21000Processed variant 22000Processed variant 23000Processed variant 24000Processed variant 25000Processed variant 26000Processed variant 27000Processed variant 28000Processed variant 29000Processed variant 30000Processed variant 31000Processed variant 32000Processed variant 33000Processed variant 34000Processed variant 35000Processed variant 36000Processed variant 37000Processed variant 38000Processed variant 39000Processed variant 40000Processed variant 41000Processed variant 42000Processed variant 43000Processed variant 44000Processed variant 45000Processed variant 46000Processed variant 47000Processed variant 48000Processed variant 49000Processed variant 50000Processed variant 51000Processed variant 52000Processed variant 53000Processed variant 54000Processed variant 55000Processed variant 56000Processed variant 57000Processed variant 58000Processed variant 59000Processed variant 60000Processed variant 61000Processed variant 62000Processed variant 63000Processed variant 64000Processed variant 65000Processed variant 66000Processed variant 67000Processed variant 68000Processed variant 69000Processed variant 70000Processed variant 71000Processed variant 72000Processed variant 73000Processed variant 74000Processed variant 75000Processed variant 76000Processed variant 77000Processed variant 78000Processed variant 79000Processed variant 80000Processed variant 81000Processed variant 82000Processed variant 83000Processed variant 84000Processed variant 85000Processed variant 86000Processed variant 87000Processed variant 88000Processed variant 89000Processed variant 90000Processed variant 91000Processed variant 92000Processed variant 93000Processed variant 94000Processed variant 95000Processed variant 96000Processed variant 97000Processed variant 98000Processed variant 99000Processed variant 100000Processed variant 101000Processed variant 102000Processed variant 103000Processed variant 104000Processed variant 105000Processed variant 106000Processed variant 107000Processed variant 108000Processed variant 109000Processed variant 110000Processed variant 111000Processed variant 112000Processed variant 113000Processed variant 114000Processed variant 115000Processed variant 116000Processed variant 117000Processed variant 118000Processed variant 119000Processed variant 120000Processed variant 121000Processed variant 122000Processed variant 123000Processed variant 124000Processed variant 125000Processed variant 126000Processed variant 127000Processed variant 128000Processed variant 129000Processed variant 130000Processed variant 131000Processed variant 132000Processed variant 133000Processed variant 134000Processed variant 135000Processed variant 136000Processed variant 137000Processed variant 138000Processed variant 139000Processed variant 140000Processed variant 141000Processed variant 142000Processed variant 143000Processed variant 144000Processed variant 145000Processed variant 146000Processed variant 147000Processed variant 148000Processed variant 149000Processed variant 150000Processed variant 151000Processed variant 152000Processed variant 153000Processed variant 154000Processed variant 155000Processed variant 156000Processed variant 157000Processed variant 158000Processed variant 159000Processed variant 160000Processed variant 161000Processed variant 162000Processed variant 163000Processed variant 164000Processed variant 165000Processed variant 166000Processed variant 167000Processed variant 168000Processed variant 169000Processed variant 170000Processed variant 171000Processed variant 172000Processed variant 173000Processed variant 174000Processed variant 175000Processed variant 176000Processed variant 177000Processed variant 178000Processed variant 179000Processed variant 180000Processed variant 181000Processed variant 182000Processed variant 183000Processed variant 184000Processed variant 185000Processed variant 186000Processed variant 187000Processed variant 188000Processed variant 189000Processed variant 190000Processed variant 191000Processed variant 192000Processed variant 193000Processed variant 194000Processed variant 195000Processed variant 196000Processed variant 197000Processed variant 198000Processed variant 199000Processed variant 200000Processed variant 201000Processed variant 202000Processed variant 203000Processed variant 204000Processed variant 205000Processed variant 206000Processed variant 207000Processed variant 208000Processed variant 209000Processed variant 210000Processed variant 211000Processed variant 212000Processed variant 213000Processed variant 214000Processed variant 215000Processed variant 216000Processed variant 217000Processed variant 218000Processed variant 219000Processed variant 220000Processed variant 221000Processed variant 222000Processed variant 223000Processed variant 224000Processed variant 225000Processed variant 226000Processed variant 227000Processed variant 228000Processed variant 229000Processed variant 230000Processed variant 231000Processed variant 232000Processed variant 233000Processed variant 234000Processed variant 235000Processed variant 236000Processed variant 237000Processed variant 238000Processed variant 239000Processed variant 240000Processed variant 241000Processed variant 242000Processed variant 243000Processed variant 244000Processed variant 245000Processed variant 246000Processed variant 247000Processed variant 248000Processed variant 249000Processed variant 250000Processed variant 251000Processed variant 252000Processed variant 253000Processed variant 254000Processed variant 255000Processed variant 256000Processed variant 257000Processed variant 258000Processed variant 259000Processed variant 260000Processed variant 261000Processed variant 262000Processed variant 263000Processed variant 264000Processed variant 265000Processed variant 266000Processed variant 267000Processed variant 268000Processed variant 269000Processed variant 270000Processed variant 271000Processed variant 272000Processed variant 273000Processed variant 274000Processed variant 275000Processed variant 276000Processed variant 277000Processed variant 278000Processed variant 279000Processed variant 280000Processed variant 281000Processed variant 282000Processed variant 283000Processed variant 284000Processed variant 285000Processed variant 286000Processed variant 287000Processed variant 288000Processed variant 289000Processed variant 290000Processed variant 291000Processed variant 292000Processed variant 293000Processed variant 294000Processed variant 295000Processed variant 296000Processed variant 297000Processed variant 298000Processed variant 299000Processed variant 300000Processed variant 301000Processed variant 302000Processed variant 303000Processed variant 304000Processed variant 305000Processed variant 306000Processed variant 307000Processed variant 308000Processed variant 309000Processed variant 310000Processed variant 311000Processed variant 312000Processed variant 313000Processed variant 314000Processed variant 315000Processed variant 316000Processed variant 317000Processed variant 318000Processed variant 319000Processed variant 320000Processed variant 321000Processed variant 322000Processed variant 323000Processed variant 324000Processed variant 325000Processed variant 326000Processed variant 327000Processed variant 328000Processed variant 329000Processed variant 330000Processed variant 331000Processed variant 332000Processed variant 333000Processed variant 334000Processed variant 335000Processed variant 336000Processed variant 337000Processed variant 338000Processed variant 339000Processed variant 340000Processed variant 341000Processed variant 342000Processed variant 343000Processed variant 344000Processed variant 345000Processed variant 346000Processed variant 347000Processed variant 348000Processed variant 349000Processed variant 350000Processed variant 351000Processed variant 352000Processed variant 353000Processed variant 354000Processed variant 355000Processed variant 356000Processed variant 357000Processed variant 358000Processed variant 359000Processed variant 360000Processed variant 361000Processed variant 362000Processed variant 363000Processed variant 364000Processed variant 365000Processed variant 366000Processed variant 367000Processed variant 368000Processed variant 369000Processed variant 370000Processed variant 371000Processed variant 372000Processed variant 373000Processed variant 374000Processed variant 375000Processed variant 376000Processed variant 377000Processed variant 378000Processed variant 379000Processed variant 380000Processed variant 381000Processed variant 382000Processed variant 383000Processed variant 384000Processed variant 385000Processed variant 386000Processed variant 387000Processed variant 388000Processed variant 389000Processed variant 390000Processed variant 391000Processed variant 392000Processed variant 393000Processed variant 394000Processed variant 395000Processed variant 396000Processed variant 397000Processed variant 398000Processed variant 399000Processed variant 400000Processed variant 401000Processed variant 402000Processed variant 403000Processed variant 404000Processed variant 405000Processed variant 406000Processed variant 407000Processed variant 408000Processed variant 409000Processed variant 410000Processed variant 411000Processed variant 412000Processed variant 413000Processed variant 414000Processed variant 415000Processed variant 416000Processed variant 417000Processed variant 418000Processed variant 419000Processed variant 420000Processed variant 421000Processed variant 422000Processed variant 423000Processed variant 424000Processed variant 425000Processed variant 426000Processed variant 427000Processed variant 428000Processed variant 429000Processed variant 430000Processed variant 431000Processed variant 432000Processed variant 433000Processed variant 434000Processed variant 435000Processed variant 436000Processed variant 437000Processed variant 438000Processed variant 439000Processed variant 440000Processed variant 441000Processed variant 442000Processed variant 443000Processed variant 444000Processed variant 445000Processed variant 446000Processed variant 447000Processed variant 448000Processed variant 449000Processed variant 450000Processed variant 451000Processed variant 452000Processed variant 453000Processed variant 454000Processed variant 455000Processed variant 456000Processed variant 457000Processed variant 458000Processed variant 459000Processed variant 460000Processed variant 461000Processed variant 462000Processed variant 463000Processed variant 464000Processed variant 465000Processed variant 466000Processed variant 467000Processed variant 468000Processed variant 469000Processed variant 470000Processed variant 471000Processed variant 472000Processed variant 473000Processed variant 474000Processed variant 475000Processed variant 476000Processed variant 477000Processed variant 478000Processed variant 479000Processed variant 480000Processed variant 481000Processed variant 482000Processed variant 483000Processed variant 484000Processed variant 485000Processed variant 486000Processed variant 487000Processed variant 488000Processed variant 489000Processed variant 490000Processed variant 491000Processed variant 492000Processed variant 493000Processed variant 494000Processed variant 495000Processed variant 496000Processed variant 497000Processed variant 498000Processed variant 499000Processed variant: 499848  
## All variants processed

raw.vcfEU <- nrow(vcfEU)

# Calculating the number of polymorphic sites and removing all non-biallelic sites

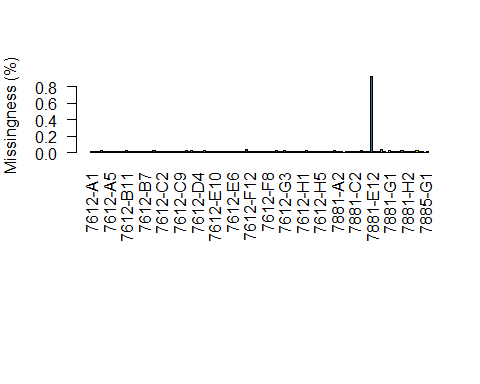
vcfNA <- vcfNA[is.polymorphic(vcfNA, na.omit = T)]  
poly.vcfNA <- nrow(vcfNA)  
vcfNA <- vcfNA[is.biallelic(vcfNA)]  
biallelic.vcfNA <- nrow(vcfNA)  
  
vcfEU <- vcfEU[is.polymorphic(vcfEU, na.omit = T)]  
poly.vcfEU <- nrow(vcfEU)  
vcfEU <- vcfEU[is.biallelic(vcfEU)]  
biallelic.vcfEU <- nrow(vcfEU)

# Backing up files

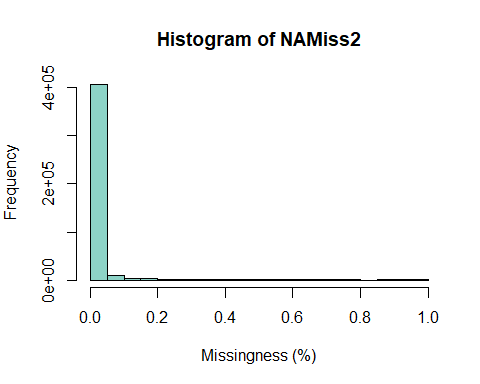
dp.vcNA <- extract.gt(vcfNA, element = "DP", as.numeric = T)  
gt.vcNA <- extract.gt(vcfNA, element = "GT", as.numeric = T)  
  
# Backup  
dp.vc.bk.na <- dp.vcNA  
gt.vc.bk.na <- gt.vcNA  
  
dp.vcEU <- extract.gt(vcfEU, element = "DP", as.numeric = T)  
gt.vcEU <- extract.gt(vcfEU, element = "GT", as.numeric = T)  
  
# Backup  
dp.vc.bk.eu <- dp.vcEU  
gt.vc.bk.eu <- gt.vcEU

# Visualizing missing data

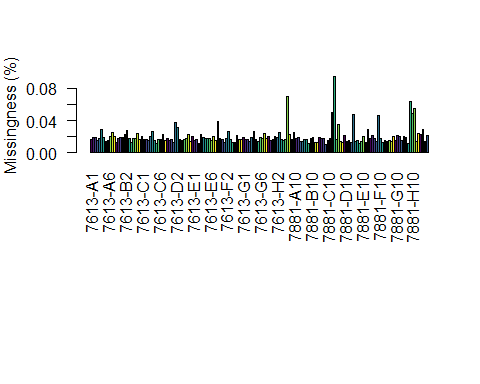
palette(viridis(n=12))  
  
NAMiss <- apply(gt.vcNA, MARGIN = 2, function(x){ sum(is.na(x)) })  
NAMiss <- NAMiss/nrow(vcfNA)  
  
par(mar = c(12,4,4,2))  
barplot(NAMiss, las = 2, col = 1:12)  
title(ylab = "Missingness (%)")



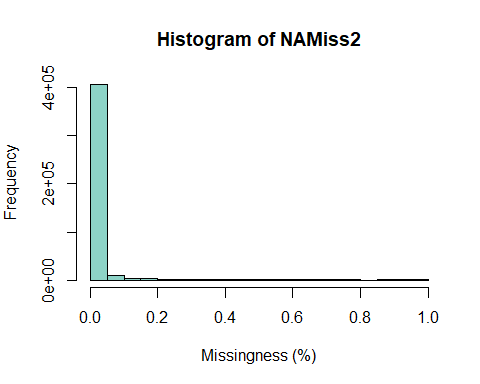
par(mar = c(5,4,4,2))  
NAMiss2 <- apply(gt.vcNA, MARGIN = 1, function(x){ sum(is.na(x)) })  
NAMiss2 <- NAMiss2/ncol(vcfNA@gt[,-1])  
  
hist(NAMiss2, col = "#8DD3C7", xlab = "Missingness (%)")



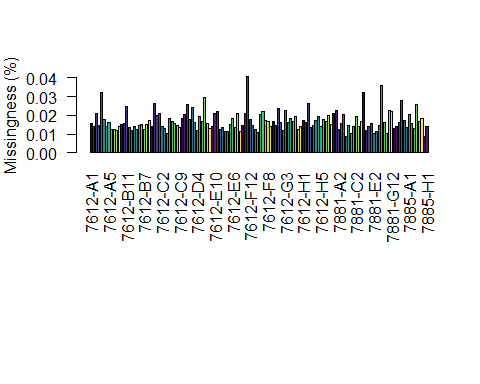
EUMiss <- apply(gt.vcEU, MARGIN = 2, function(x){ sum(is.na(x)) })  
EUMiss <- EUMiss/nrow(vcfEU)  
  
par(mar = c(12,4,4,2))  
barplot(EUMiss, las = 2, col = 1:12)  
title(ylab = "Missingness (%)")



par(mar = c(5,4,4,2))  
EUMiss2 <- apply(gt.vcEU, MARGIN = 1, function(x){ sum(is.na(x)) })  
EUMiss2 <- EUMiss2/ncol(vcfEU@gt[,-1])  
  
hist(NAMiss2, col = "#8DD3C7", xlab = "Missingness (%)")



#Removing an extreme outlier (7881-E12 has 92% missingness)  
gt.vcNA <- gt.vcNA[,-113]  
  
NAMiss <- apply(gt.vcNA, MARGIN = 2, function(x){ sum(is.na(x)) })  
NAMiss <- NAMiss/nrow(vcfNA)  
  
par(mar = c(12,4,4,2))  
barplot(NAMiss, las = 2, col = 1:12)  
title(ylab = "Missingness (%)")



## Filtering by DP (DP (1?x < x > 95%))

# Creating quantiles based on the lower 5% and the upper 95%  
sumsNA <- apply(dp.vc.bk.na, 2,   
 function (x) quantile(x, probs=c(0.05,   
 0.50,   
 0.95),   
 na.rm = T))  
  
# Add the minimum depth (14x)  
sumsNA[1,][sumsNA[1,] < 14] <- 14  
dp.all.na <- sweep(dp.vc.bk.na, MARGIN=2, FUN="-", sumsNA[1,])  
dp.vcNA[dp.all.na <= 0] <- NA  
dp.all.na <- sweep(dp.vc.bk.na, MARGIN=2, FUN="-", sumsNA[3,])  
dp.vcNA[dp.all.na > 0] <- NA  
  
# Changing GT  
vcfNA@gt[,-1][is.na(dp.vcNA)] <- NA  
  
# Creating quantiles based on the lower 5% and the upper 95%  
sumsEU <- apply(dp.vc.bk.eu, 2,   
 function (x) quantile(x, probs=c(0.05,   
 0.50,   
 0.95),   
 na.rm = T))  
  
# Add the minimum depth (12x)  
sumsEU[1,][sumsEU[1,] < 12] <- 12  
dp.all.eu <- sweep(dp.vc.bk.eu, MARGIN=2, FUN="-", sumsEU[1,])  
dp.vcEU[dp.all.eu <= 0] <- NA  
dp.all.eu <- sweep(dp.vc.bk.eu, MARGIN=2, FUN="-", sumsEU[3,])  
dp.vcEU[dp.all.eu > 0] <- NA  
  
# Changing GT  
vcfEU@gt[,-1][is.na(dp.vcEU)] <- NA

# Filtering by maximum MQ (MQ == 50)

mq.na <- extract.info(vcfNA, element = "MQ", as.numeric = T)  
  
# Creating mask  
mask.mq.na <- rep(T, nrow(vcfNA))  
  
# Filtering in the mask  
mask.mq.na[mq.na < 50] <- F  
  
# Regions removed by mask  
biallelic.vcfNA - sum(mask.mq.na)

## [1] 103597

mq.eu <- extract.info(vcfEU, element = "MQ", as.numeric = T)  
  
# Creating mask  
mask.mq.eu <- rep(T, nrow(vcfEU))  
  
# Filtering in the mask  
mask.mq.eu[mq.eu < 50] <- F  
  
# Regions removed by mask  
biallelic.vcfEU - sum(mask.mq.eu)

## [1] 100496

# Filtering on MAF (5% minor allele freq)

maf.tresh.na <- 7/(ncol(vcfNA@gt[,-1]))  
cat("MAF threshold:", maf.tresh.na, "\n")

## MAF threshold: 0.05185185

# Creating mask  
mask.maf.na <- rep(T, nrow(vcfNA))  
  
# Extracting GT and calculating MAF  
class(gt.vcNA) <- 'numeric'  
mask.maf.na <- apply(gt.vcNA, 1,   
 function (x) min(table(x)))/ncol(vcfNA@gt) >= maf.tresh.na  
  
# Regions removed by mask  
biallelic.vcfNA - sum(mask.maf.na)

## [1] 143779

maf.tresh.eu <- 8/(ncol(vcfEU@gt[,-1]))  
cat("MAF threshold:", maf.tresh.eu, "\n")

## MAF threshold: 0.05

# Creating mask  
mask.maf.eu <- rep(T, nrow(vcfEU))  
  
# Extracting GT and calculating MAF  
class(gt.vcEU) <- 'numeric'  
mask.maf.eu <- apply(gt.vcEU, 1,   
 function (x) min(table(x)))/ncol(vcfEU@gt) >= maf.tresh.eu  
  
# Regions removed by mask  
biallelic.vcfEU - sum(mask.maf.eu)

## [1] 53456

# Filtering by missing data:

# Creating mask  
mask.miss.na <- rep(T, nrow(vcfNA))  
  
# Extracting GT and calculating missingness  
gt.vcNA <- extract.gt(vcfNA, element = "GT", as.numeric = T)  
mask.miss.na <- apply(gt.vcNA, 1,   
 function (x) sum(is.na(x),  
 na.rm = TRUE)/ncol(gt.vcNA)) <= 0.10  
  
# Regions removed by mask  
biallelic.vcfNA - sum(mask.miss.na)

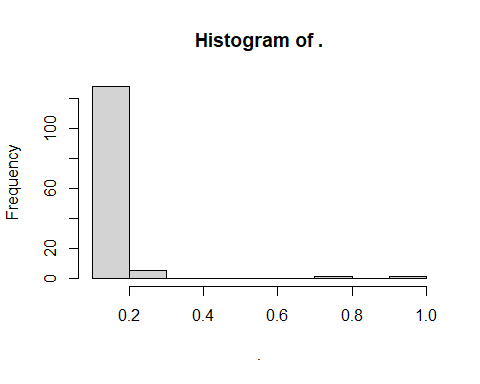
## [1] 132845

# Creating mask  
mask.miss.eu <- rep(T, nrow(vcfEU))  
  
# Extracting GT and calculating missingness  
gt.vcEU <- extract.gt(vcfEU, element = "GT", as.numeric = T)  
mask.miss.eu <- apply(gt.vcEU, 1,   
 function (x) sum(is.na(x),  
 na.rm = TRUE)/ncol(gt.vcEU)) <= 0.10  
  
# Regions removed by mask  
biallelic.vcfEU - sum(mask.miss.eu)

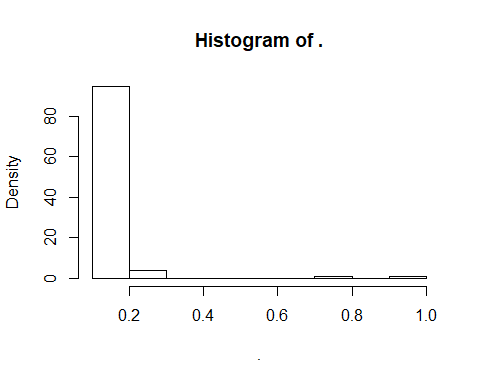
## [1] 98119

## Missing DP per sample (Missing more than ??% data)

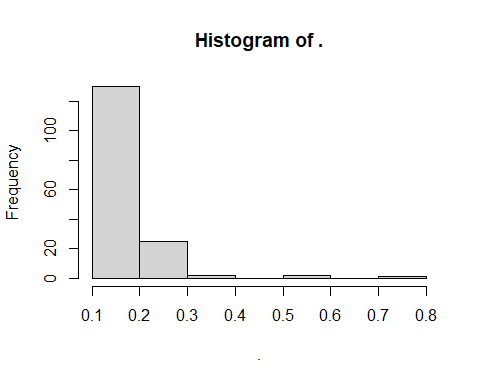
h.na <- apply(dp.vcNA, 2,   
 function (x) is.na(x) %>% sum/nrow(dp.vcNA)) %>% hist()



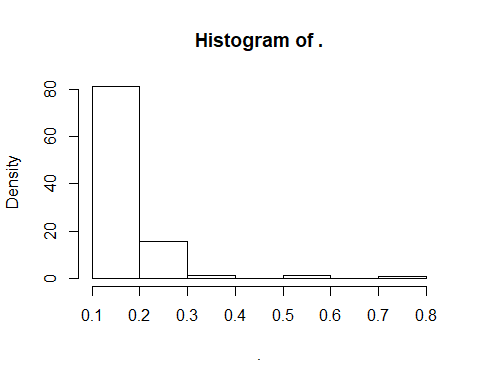
h.na$density = h.na$counts/sum(h.na$counts)\*100  
plot(h.na,freq=FALSE)



h.eu <- apply(dp.vcEU, 2,   
 function (x) is.na(x) %>% sum/nrow(dp.vcEU)) %>% hist()



h.eu$density = h.eu$counts/sum(h.eu$counts)\*100  
plot(h.eu,freq=FALSE)



missing.na <- colnames(dp.vcNA)[apply(dp.vcNA, 2,   
 function (x) is.na(x)  
 %>%  
 sum/nrow(dp.vcNA)) > 0.2]  
vcfNA@gt <- vcfNA@gt[,!colnames(vcfNA@gt) %in% missing.na]  
mask.miss.na <- rep(T, nrow(vcfNA))  
gt.vcNA <- extract.gt(vcfNA, element = "GT", as.numeric = T)  
mask.miss.na <- apply(gt.vcNA, 1,   
 function (x) sum(is.na(x))/ncol(gt.vcNA)) <= 0.00  
  
missing.eu <- colnames(dp.vcEU)[apply(dp.vcEU, 2,   
 function (x) is.na(x)  
 %>%  
 sum/nrow(dp.vcEU)) > 0.2]  
vcfEU@gt <- vcfEU@gt[,!colnames(vcfEU@gt) %in% missing.eu]  
mask.miss.eu <- rep(T, nrow(vcfEU))  
gt.vcEU <- extract.gt(vcfEU, element = "GT", as.numeric = T)  
mask.miss.eu <- apply(gt.vcEU, 1,   
 function (x) sum(is.na(x))/ncol(gt.vcEU)) <= 0.00

# Processing masks

mask.vcf.na <- cbind(mask.mq.na, mask.maf.na, mask.miss.na)  
filtered.vcf.na <- vcfNA[apply(mask.vcf.na, 1, sum) == 3,]  
filtered.vcf.na

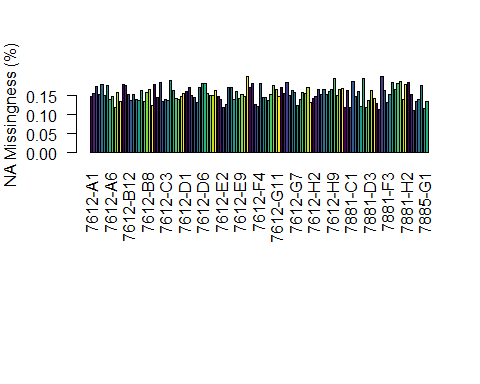
## \*\*\*\*\* Object of Class vcfR \*\*\*\*\*  
## 128 samples  
## 28 CHROMs  
## 50,693 variants  
## Object size: 479.3 Mb  
## 0 percent missing data  
## \*\*\*\*\* \*\*\*\*\* \*\*\*\*\*

mask.vcf.eu <- cbind(mask.mq.eu, mask.maf.eu, mask.miss.eu)  
filtered.vcf.eu <- vcfEU[apply(mask.vcf.eu, 1, sum) == 3,]  
filtered.vcf.eu

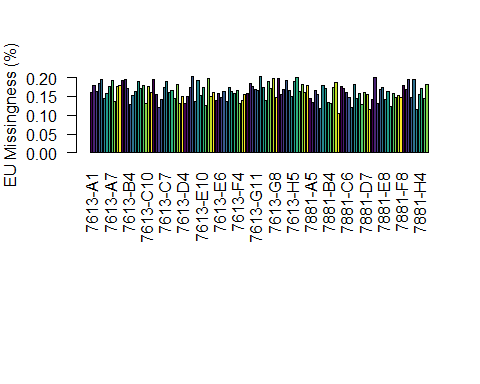
## \*\*\*\*\* Object of Class vcfR \*\*\*\*\*  
## 130 samples  
## 27 CHROMs  
## 63,736 variants  
## Object size: 522.2 Mb  
## 0 percent missing data  
## \*\*\*\*\* \*\*\*\*\* \*\*\*\*\*

# Revisualizing missingness

gt.vcNA <- extract.gt(vcfNA, element = "GT", as.numeric = T)  
NAMiss <- apply(gt.vcNA, MARGIN = 2, function(x){ sum(is.na(x)) })  
NAMiss <- NAMiss/nrow(vcfNA)  
  
par(mar = c(12,4,4,2))  
barplot(NAMiss, las = 2, col = 1:12)  
title(ylab = "NA Missingness (%)")



gt.vcEU <- extract.gt(vcfEU, element = "GT", as.numeric = T)  
EUMiss <- apply(gt.vcEU, MARGIN = 2, function(x){ sum(is.na(x)) })  
EUMiss <- EUMiss/nrow(vcfEU)  
  
par(mar = c(12,4,4,2))  
barplot(EUMiss, las = 2, col = 1:12)  
title(ylab = "EU Missingness (%)")

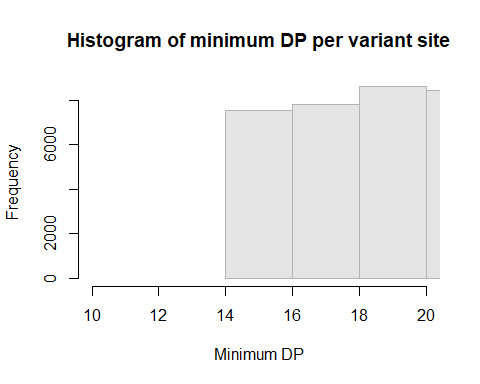


# Recalculating the number of polymorphic sites and removing all non-biallelic sites

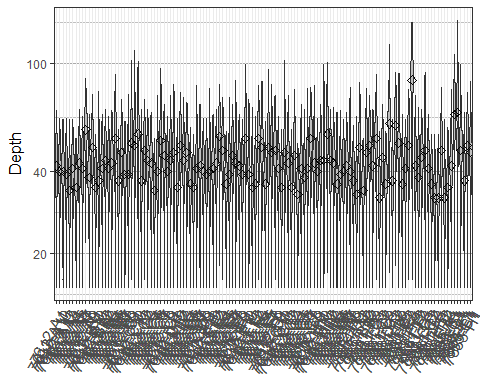
filtered.vcf.na <- filtered.vcf.na[is.polymorphic(filtered.vcf.na, na.omit = T)]  
poly.vcfNA <- nrow(filtered.vcf.na)  
vcfNA <- filtered.vcf.na[is.biallelic(filtered.vcf.na)]  
biallelic.vcfNA <- nrow(filtered.vcf.na)  
write.vcf(filtered.vcf.na, file = "FinalNA.vcf.gz", mask = F)  
  
filtered.vcf.eu <- filtered.vcf.eu[is.polymorphic(filtered.vcf.eu, na.omit = T)]  
poly.vcfEU <- nrow(filtered.vcf.eu)  
vcfEU <- filtered.vcf.eu[is.biallelic(filtered.vcf.eu)]  
biallelic.vcfEU <- nrow(filtered.vcf.eu)  
write.vcf(filtered.vcf.eu, file = "FinalEU.vcf.gz", mask = F)

# Did the new filtering by DP work?

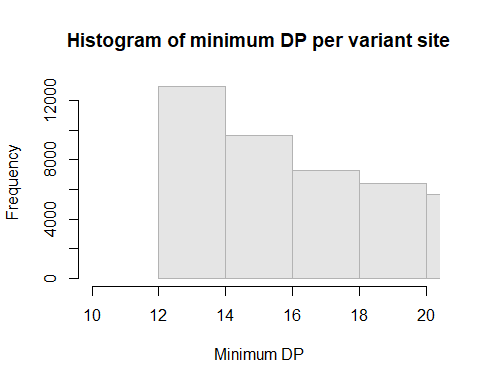
# DP from the new filtered data:  
dp.filtered.na <- extract.gt(filtered.vcf.na, element = "DP", as.numeric = T)  
apply(dp.filtered.na, 1, min) %>%   
 hist(xlab="Minimum DP",   
 main="Histogram of minimum DP per variant site",   
 border="grey70", col="grey90", breaks=10, xlim=c(10,20))



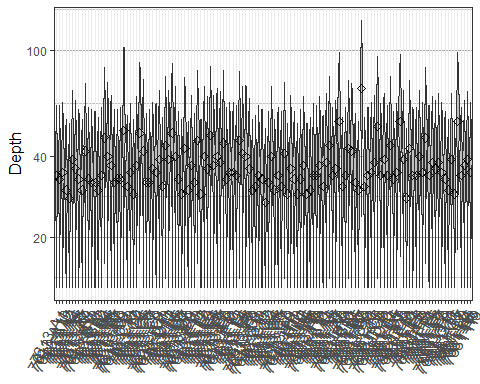
dpfNA <- melt(dp.filtered.na, varnames=c('Index', 'Sample'),   
 value.name = 'Depth', na.rm=TRUE)  
  
p <- ggplot(dpfNA, aes(x=Sample, y=Depth))+   
 geom\_violin(fill="#C0C0C0", adjust=1.0,  
 scale = "count", trim=TRUE)+   
 theme\_bw()+  
 theme(axis.title.x = element\_blank(),   
 axis.text.x = element\_text(angle = 60,   
 hjust = 1,   
 size=12),  
 axis.title.y = element\_text(size=12),  
 panel.grid.major.y=element\_line(color = "#A9A9A9",   
 size=0.6),  
 panel.grid.minor.y=element\_line(color = "#C0C0C0",   
 size=0.2))+   
 scale\_y\_continuous(trans=scales::log2\_trans(),   
 breaks=c(1, 10, 20, 40, 100))+  
 stat\_summary(fun=median, geom="point", shape=23, size=2)  
p



# DP from the new filtered data:  
dp.filtered.eu <- extract.gt(filtered.vcf.eu, element = "DP", as.numeric = T)  
apply(dp.filtered.eu, 1, min) %>%   
 hist(xlab="Minimum DP",   
 main="Histogram of minimum DP per variant site",   
 border="grey70", col="grey90", breaks=10, xlim=c(10,20))



dpfEU <- melt(dp.filtered.eu, varnames=c('Index', 'Sample'),   
 value.name = 'Depth', na.rm=TRUE)  
  
q <- ggplot(dpfEU, aes(x=Sample, y=Depth))+   
 geom\_violin(fill="#C0C0C0", adjust=1.0,  
 scale = "count", trim=TRUE)+   
 theme\_bw()+  
 theme(axis.title.x = element\_blank(),   
 axis.text.x = element\_text(angle = 60,   
 hjust = 1,   
 size=12),  
 axis.title.y = element\_text(size=12),  
 panel.grid.major.y=element\_line(color = "#A9A9A9",  
 size=0.6),  
 panel.grid.minor.y=element\_line(color = "#C0C0C0",  
 size=0.2))+   
 scale\_y\_continuous(trans=scales::log2\_trans(),   
 breaks=c(1, 10, 20, 40, 100))+  
 stat\_summary(fun=median, geom="point", shape=23, size=2)  
q



# Testing for heterozygosity

gt.vc.het.na <- extract.gt(vcfNA, element = "GT", as.numeric = F)  
het.na <- is.het(gt.vc.het.na)  
het.sums.na <- rowSums(het.na)  
#view(het.sums.na)  
  
gt.vc.het.eu <- extract.gt(vcfEU, element = "GT", as.numeric = F)  
het.eu <- is.het(gt.vc.het.eu)  
het.sums.eu <- rowSums(het.eu)  
view(het.sums.eu)