# Proves d'anàlisi

*Ruth Gómez Graciani*

## Test de qualitat per l'ajustament de la resolució

"Event de recombinació": rang proporcionat en el paper de Bell et al. 2020 (~400kb de llarg).

"Entrecreuament": posició dins de l'event de recombinació on hi hauria hagut l'entrecreuament.

En el paper original es considera que els entrecreuaments han tingut lloc al centre dels events de recombinació. Aquesta situacio, en especial tenint en compte la mida dels events de recombinació donats, és poc realista. Per això nostraltres provarem de calcular la taxa de recombinacio en cada una de les finestres sumant el valor de les parts proporcionals dels events de recombinacio que hi solapen, de forma que es dona mes credibilitat als events de recombinacio d'alta resolucio. Aixo ens permetria, a mes, tenir mapes elaborats amb finestres mes petites sense que el valor de moltes d'aquestes finestres sigui 0 tot i que solapin amb events de recombinacio i per tant existia la possibilitat de que s'hagues efectuat un entrecreuament dins els limits de la finestra en qüestió.

Per conèixer els efectes de l'augment de la resolució sobre les estimes de recombinació i assegurar-nos que no causem esbiaixos importants, es fa uncontrol de qualitat comparant els dos metodes i tambe comparant els resultats d'agafar mides de finestra mes grans o mes petites (500kb, 100kb, 50kb i 20kb). Tot i que a nivell local podria haver-hi diferències puntuals, en termes globals la nostra metodologia hauria de donar resultats semblants a la del paper.

Per saber si el sistema de mesura està afectant al mapa de recombinació, podem comprovar si hi ha una correlació entre els mapes calculats d'una manera i l'altra (Figura 1). Els resultats mostren una correlació entre moderada i forta, estadísticament significativa, pero en els gràfics es pot veure com algunes de les taxes de recombinació més altes en el paper original són bastant inferiors amb el nostre mètode, independentment de la mida de finestra. Això és degut a que alguns d'aquests punts probablement solapen amb molts events de recombinació representats amb intervals de confiança molt grans i, per tant, la probabilitat real de que l'entrecreuament hagi tingut lloc específicament a la finestra en qüestió és menor que l'estimat simplement tenint en compte el punt central de l'event de recombinació.

Aquest efecte a nivell local sobre les taxes de recombinació és esperable i desitjable, ja que el que vol dir que les nostres mesures estan modulant falsos pics alhora que augmenta la resolució del mapa. En cas que realment estigui funcionant correctament, aquesta modulació local no afectarà a la mitjana de les finestres, és a dir, que el valor de les finestres de 500kb hauria de ser molt semblant a la mitjana de les finestres de 100, 50 o 20 kb que hi solapen. Tal com es pot veure a la Figura 2, això s'acompleix, i hi ha una correlació perfecta entre les mitjanes d'una opció i els valors de l'altra, per tant, podem concloure que alterar la mida de finestre per augmentar la resolucio no introdueix esbiaixos importants.

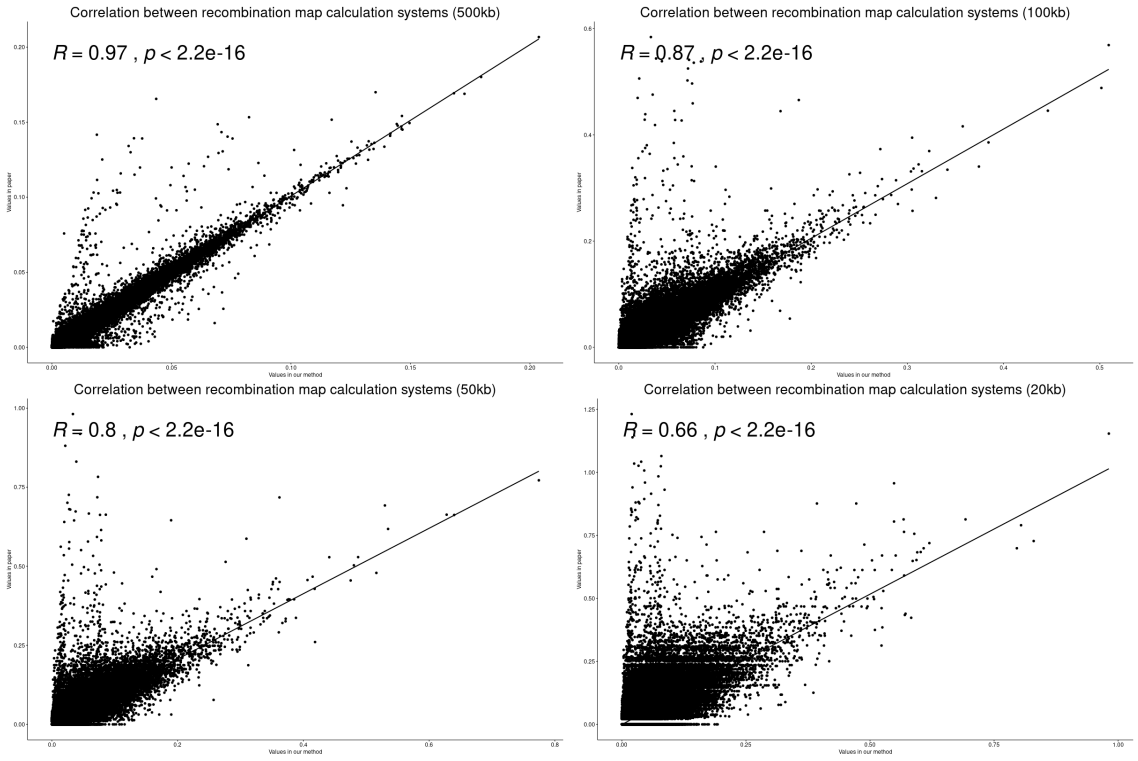


Figure 1: Correlació entre mapes de recombinacio calculats amb diferents metodologies. Per comprovar l'efecte del sistema de mesura, s'ha calculat la correlacio entre les taxes de recombinacio del mapa calculat mitjançant el metode nou (eix de les x) i del mapa calculat com el paper original (eix de les y), per mides de finestra de A=500kb, B=100kb, C=50kb, D=20kb.

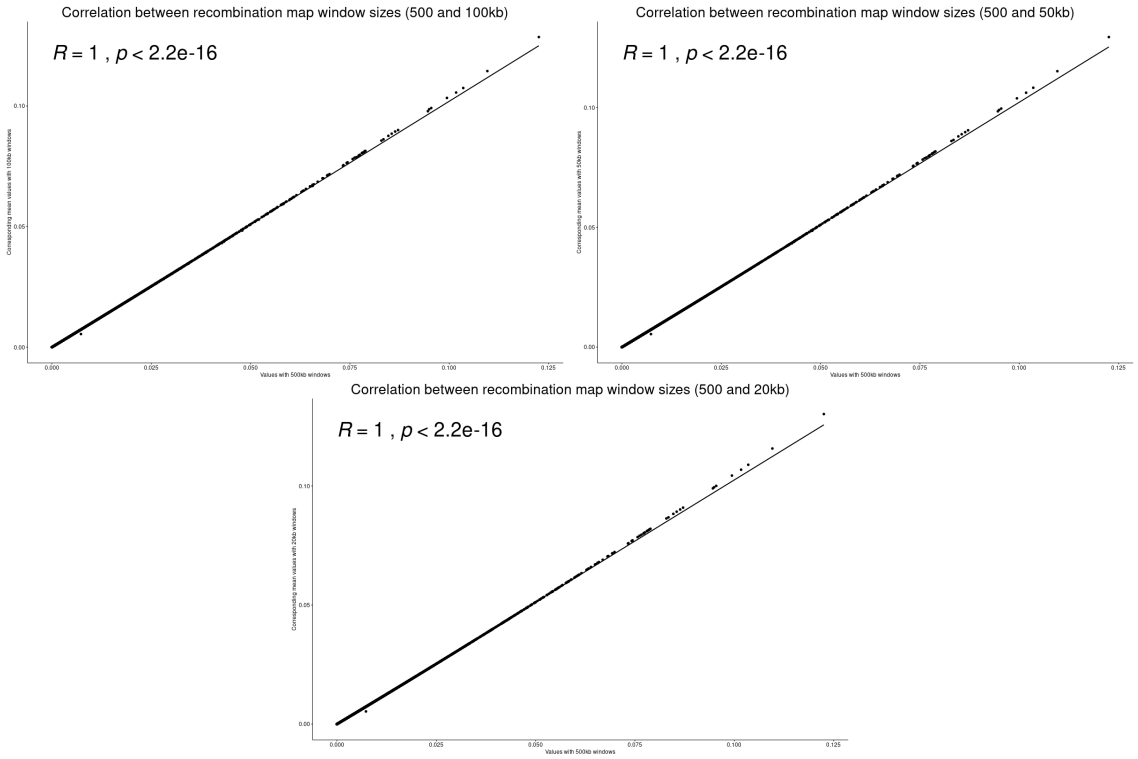


Figure 2: Correlacio entre mapes de recombinacio calculats a partir de diferents mides de finestra. Per comprovar l'efecte del sistema de mesura, s'ha calculat la correlacio entre els diferents mapes calculats mitjançant el metode nou. Per cada finestra de 500kb (eix de les x), s'ha calculat la taxa de recombinacio mitjana de les finestres de A=100kb, B=50kb, C=20kb que hi solapen.

# Mostreig de les dades de recombinació

Les inversions s'han dividit en 3 grups diferents en funció de la seva mida per analitzar-les amb paràmetres més adients. Es considerarien inversions petites aquelles menors de 5kb, mitjanes aquelles entre 5 i 50kb i grans aquelles majors de 50kb. La mida final dels grups dependrà de quantes de les inversions tenen genotips de prou qualitat com per ser analitzades i de la possibilitat d'adaptar la resolució del mapa de recombinació a la mida de finestra establerta per cada grup.

De forma general, la regió al voltant de la inversió es va dividir en "in" (zona interna de la inversió, que exclou els breakpoints), "buffer" (zona immediatament colindant a la inversió, que inclou l'interval de confiança dels breakpoints o bé els inverted repeats, el que sigui més gran, que serà d'una mida mínima de 20kb a banda i banda), i "out" (finestres de mida fixa més enllà del buffer). La mida de la regió "out" a banda i banda seria de 250kb per les inversions petites, 500kb per les mitjanes i 1Mb per les grans, dividides en finestres de 50kb.

## Filtratge del mostreig

En principi, es descartarien aquelles mostres on s'observi que no hi ha canvis al llarg de tota la regió analitzada per una inversió i individu concrets. Això sol venir causat per una baixa resolució de les mesures de recombinació locals, és a dir, que tots els events que solapen amb tota la regió de la inversió són més grans que aquesta.

Falta per contrastar aquests resultats amb dades de coverage local, ja que s'espera que menys SNPs donin lloc a events de recombinació de mida més gran i per tant menys possibilitats de millorar la resolució amb el nostre mètode de càlcul de taxes de recombinació.

Les dades de coverage local també serien útils a l'hora de diferenciar si en una regió no s'han detectat events de recombinació per falta de coverage o per una taxa de recombinació realment baixa.

Aquest seria el nombre de dades (cada una representant una inversio en un individu concret) que es perdrien en cas d'aplicar els filtres. De moment les proves i els anàlisis s'han fet amb totes les dades, a l'espera de decidir que fem en cada cas.

Taula 1. Filtratge de dades. Cada columna és un grup d'inversions segons la seva mida: G1 fins a 5kb, G2 de 5 a 50kb i G3 més de 50kb. La fila Excluded es aquella on totes les finestres per una inversio en un individu concret tenen exactament la mateixa taxa de recombinacio i per tant no podem diferenciar entre diferents regions. Aixo afectaria nomes a aquells analisis on es comparin diferents tipus de finestra dins de la mateixa inversio. La fila Recombination Rate 0 (RR0) es aquella on la taxa de recombinacio ha estat 0 en totes les finestres associades a la inversio i els seus voltants. El total es la suma de les dues files.

```
##          G1 G2 G3
## Excluded 44  2  0
## RR0      44 11  8
## Total    88 13  8
```

## Normalització del mostreig

Per poder comparar entre idividus i inversions, es va dir de normalitzar les taxes de recombinacio en funcio de l'individu i/o el cromosoma, en funcio del que es volgues fer. He fet servir el mapa de recombinació elaborat amb el nostre metode i amb finestres de 500kb com a referència per calcular les mitjanes que necessitem, ja que es la opcio computacionalment menys pesada i com ja hem demostrat abans, la mitjana sortiria la mateixa independentment de la mida de finestra triada per aquesta tasca.

En la Figura 3 no es noten diferencies apreciables quan s'aplica un o altre metode de normalitzacio, pero observant atentament la comaparativa nomes amb un subset de les dades (Figura 4) es pot veure com s'introdueixen petits canvis.

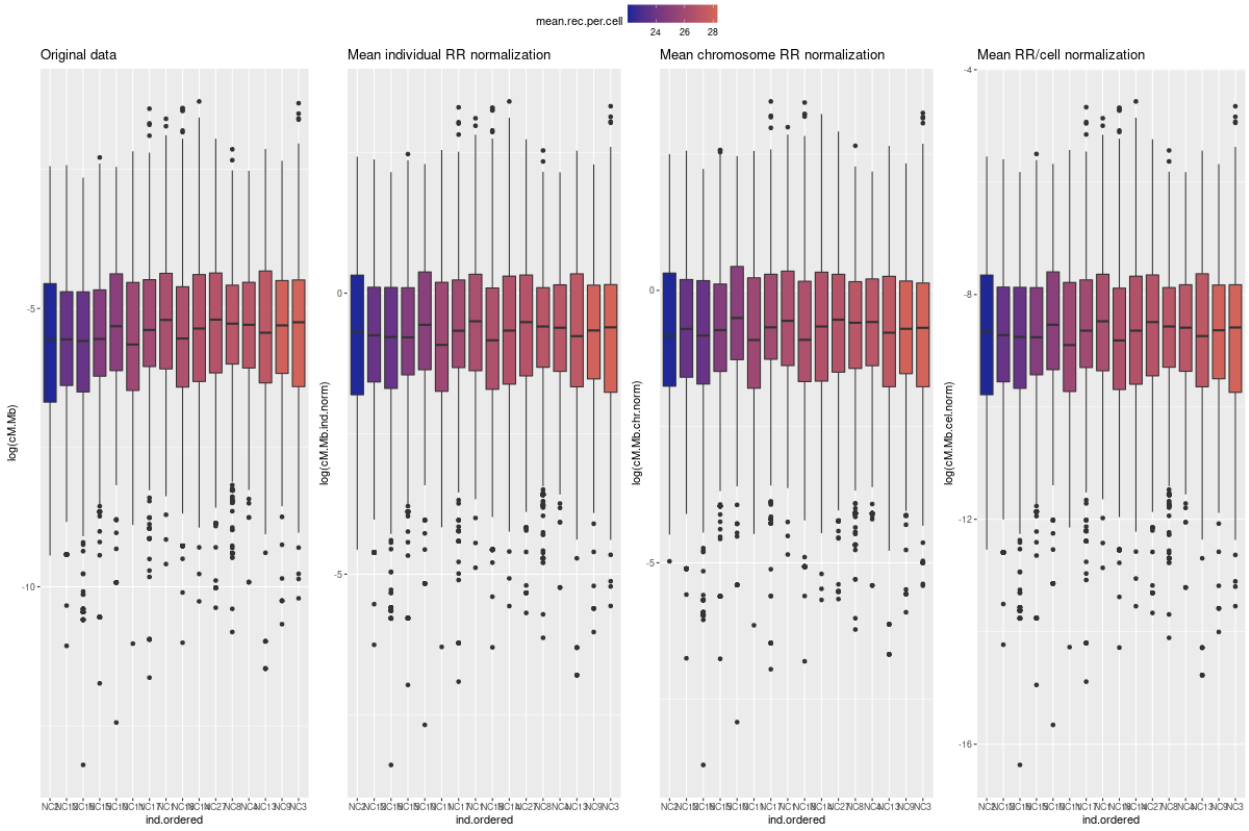


Figure 3: Efecte de diferents normalitzacions sobre les dades. Cada boxplot representa la distribucio de les taxes de recombinació al llarg del genoma per un sol individu. Els individus estan endreçants i colorejats segons el seu nivell de recombinacio global, calculat com la mitjana d'entrecreuaments per cel · lula. Depenent de la mesura que es faci servir, el eix de les y (mostrat aqui en escala logarítmica en base 10) canvia, pero la distribució de les dades es la mateixa.

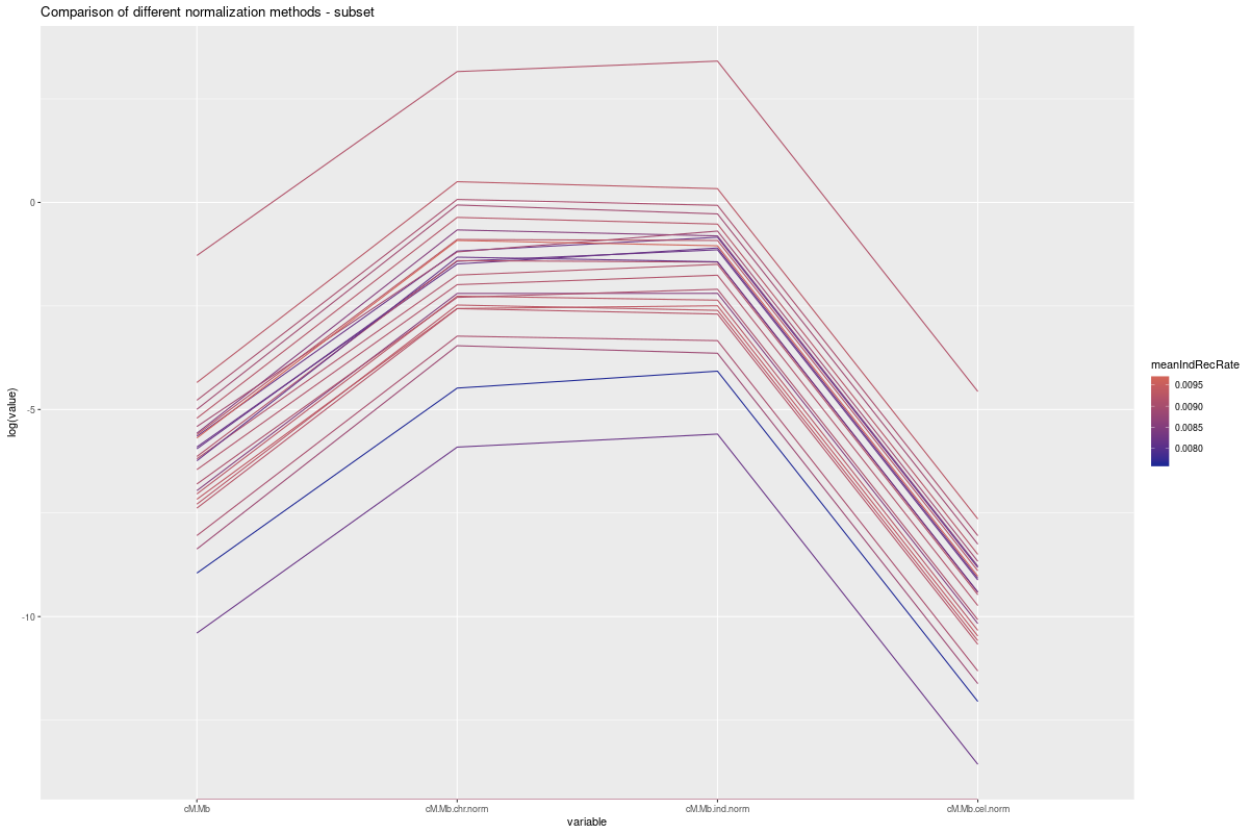


Figure 4: Efecte de diferents normalitzacions sobre les dades en una petita mostra. Cada linia representa l'evolucio d'una mateixa mesura de taxa de recombinacio per un individu i una finestra concrets. L'eix de les y esta en escala logaritmica en base 10 i el color de la linia indica la taxa de recombinacio mitjana d'aquell individu. Els valors son, d'esquerra a dreta, mesura original, mesura corregida per la taxa de recombinació mitjana del cromosoma concret, mesura corregida per la taxa de recombinació mitjana de l'individu i mesura corregida pels entrecreuaments mitjans per cel · lula en cada individu. Els canvis de la primera a la segona mesura, i de la segona a la tercera indica que la normalitzacio, tot i que subtilment, corregeix alguns esbiaixos. De la tercera a la quarta mesura les linies son paraleles perque la normalitzacio esta corregin en els dos casos per la taxa de recombinacio de l'individu.

En general, els canvis observats son molts subtils. Aixo pot ser degut a la gran similitud que hi ha entre els diferents factors de normalització, tal com es veu en els seguents resums:

```
load("Rdata/normInds.Rdata")

# Summary mean recombination events per cell
summary(unique(recRate.outwins$mean.rec.per.cell))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.19   24.94   26.40   25.92   26.98   28.13
```

```
# Summary mean chromosomal recombination rates for each chromosome and individual
summary(chr.normalization$meanChrRecRate)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.005878 0.008103 0.008933 0.009782 0.011144 0.017708
```

```
# Summary mean recombination rates for each individual
summary(ind.normalization$meanIndRecRate)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.007631 0.008792 0.009114 0.008975 0.009303 0.009652
```

Finalment, de fet, aquestes normalitzacions en concret no s'han fet servir, sino que s'ha optat per calcular quin percentil ocupa cada mesura dins de la distribució de dades per la que es vulgui corregir.

# Anàlisi estadístic

En aquesta prova hiha 64 inversions i 16 individus. La mesura que s'esta comparant es el percentil que ocupa la taxa de recombinacio en una finestra concreta dins del context del mapa de recombinacio general del seu individu corresponent. Aixo estaria per tant corregint per individu.

## Resum general

*Fent servir la regio de dins: Figura 5a* 1. A primera vista, dins d'una inversió hi pot haver qualsevol taxa de recombinació en relació amb el genoma. 2. Dit això, sembla que les inversions del grup 3 tenen taxes mes baixes en general, arribant com a maxim al ~90% major. Aixo es probable que sigui mes aviat un efecte de la mida de la finestra, al ser una mitjana d'una zona mes gran, es suavitza l'efecte. Una altra opcio es que realment hi hagi una mida a partir de la qual les taxes baixen, ja uqe hi ha una especie de clusteritzacio en les inversions del G2. 3. Una tendencia que sembla bastant generalitzada es que els heterozigots recombinen menys que els estandard. Podria ser que els invertits tinguin una taxa intermitja, pero no hi ha dades suficients de invertits en G2 i G3 com per afirmar que aixo passa, i de fet en aquests grups el nombre de inversions que finalment hem pogut analitzar es bastant justet.

*Fent servir la regio de fora: Figura 5b* 1- Sembla que aqui les taxes de recombinacio no arriben a valors tan extrems com en el cas anterior, segurament perque aqui cada punt es la mitjana de totes les finestres de la regio de fora (5, 10 i 20 finestres de 50kb pels grups G1, G2 i G3 respectivament). 2- Les tendencies que es veuen en comparar entre diferents genotips sembla que son les mateixes que fent servir la regio de dins, pero mes moderades. 3- A la regio de fora sembla haver-hi una petita tendencia a presentar valors de recombinacio majors que a la regio de dins.

## Relacio amb la mida de la inversio

S'ha fet una bateria de tests per mirar el valor de la correlacio i el seu p-value entre la mitjana dels percentils per les finestres fora de la inversio i la mida de la inversio. Fent servir les finestres de fora s'espera poder comparar, en el context del genoma d'un individu, si les inversions apareixen en regions de mes alta o baixa recombinacio.

El grup 2 presenta una forta correlacio negativa amb p-values significatius, pero estem parlant de nomes 4 inversions. Hi ha una correlacio moderada i sembla que significativa per les inversions grans quan es miren els individus heterozigots explusivament, peor altre cop aquest es un dels grups amb menys dades. Al mirar el panorama general no hi ha res remarcable, crec.

```
## $correl
##                G1         G2         G3         ALL
## STD  -0.12335999 -0.8383371 0.08268125 -0.06570374
## INV   0.14143364 -0.9831395 0.08602991 -0.25032306
## HOMO -0.08752683 -0.8510914 0.11179665 -0.07053648
## HET  -0.03963588 -0.9072322 0.71692255 -0.08916632
## ALL  -0.07421016 -0.8501085 0.19668778 -0.06750310
##
## $pval
##                G1           G2           G3         ALL
## STD  0.005646856 6.691661e-13 0.3796849919 0.091191453
## INV  0.129920926 1.686051e-02 0.8544909530 0.004533438
## HOMO 0.029581125 9.516230e-15 0.2202100752 0.047632175
## HET  0.579296251 3.044875e-06 0.0001737664 0.173096834
## ALL  0.034044878 6.370121e-19 0.0181368882 0.030778475
```

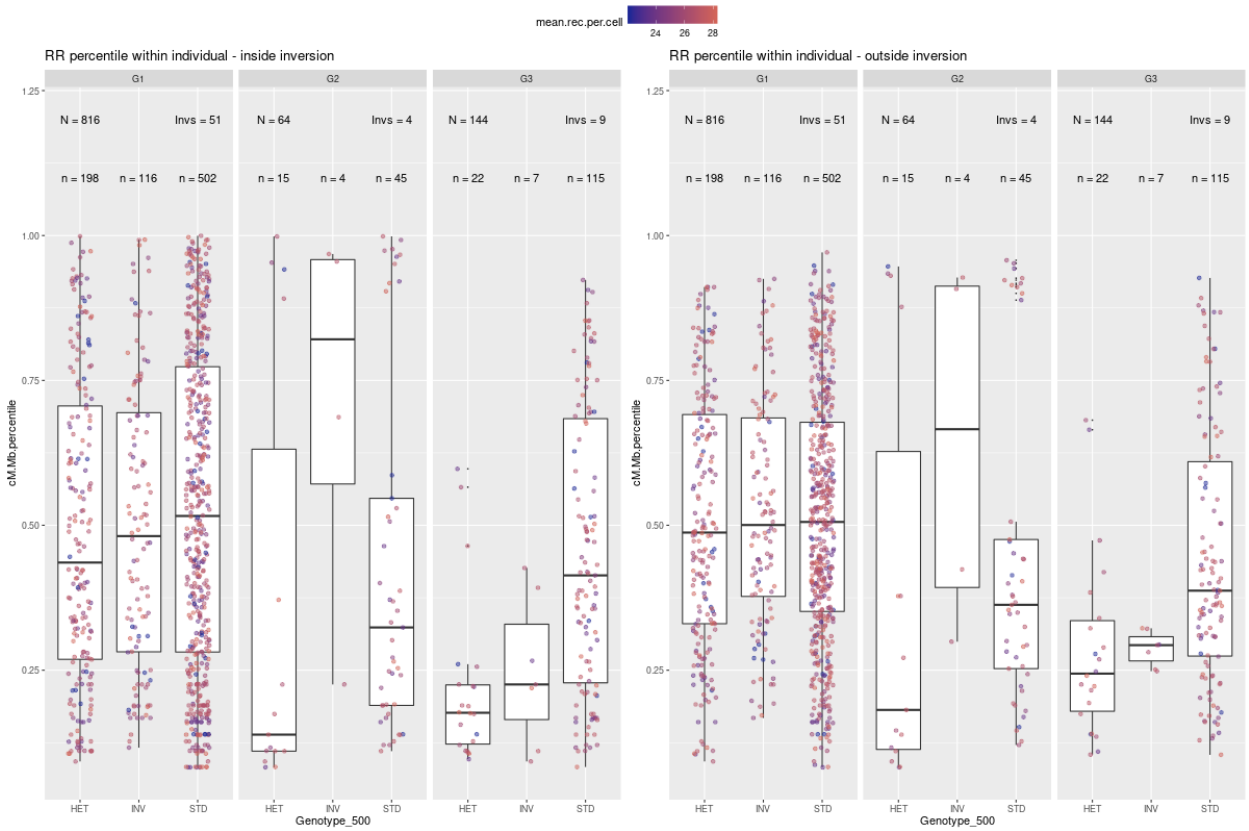


Figure 5: Distribucio general de les taxes de recombinacio dins i fora de la inversio. Les dades dins de la inversio s'han obtingut del valor de la finestra "in" per cada inversio. Les dades de fora de la inversio son la mitjana del resultat de les diferents finestres "out" de cada inversio. Cada punt representa el percentil que ocupa la taxa de recombinació per una inversio i individu concret en la regio especificada en la distribucio del mapa de recombinacio de l'individu. S'han agrupat les dades per mida de inversio (grups 1, 2, i 3) i per genotip de l'individu en qüestio. Cada punt esta colorejat segons la taxa de recombinacio mitjana d'aquell individu.

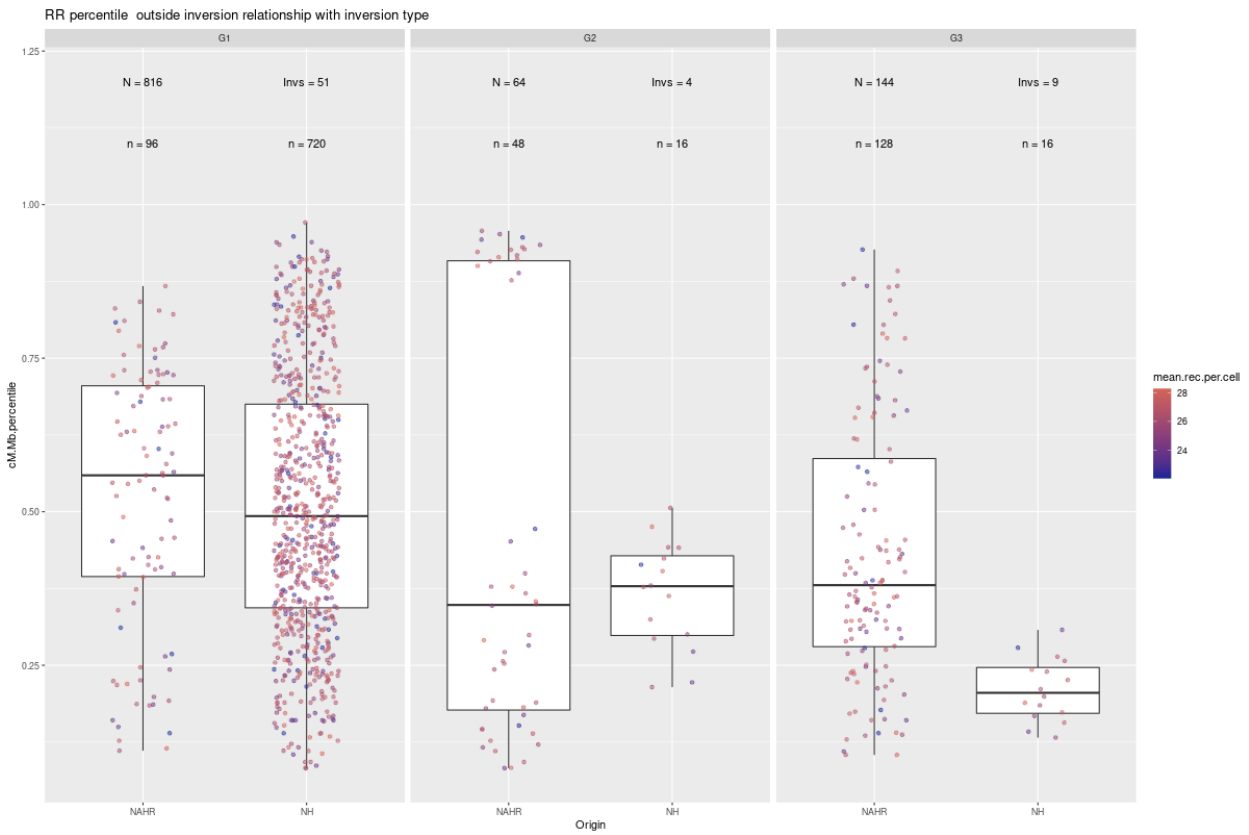


Figure 6: Taxes de recombinacio fora de la inversio en funció del mencanisme de generació. Les dades son la mitjana del resultat de les diferents finestres "out" de cada inversio per un individu determinat. Cada punt representa el percentil que ocupa la taxa de recombinació per una inversio i individu concret en la regio especificada en la distribucio del mapa de recombinacio de l'individu. S'han agrupat les dades pel mecanisme de generació i pel genotip de l'individu per la inversió en qüestio. Cada punt esta colorejat segons la taxa de recombinacio mitjana d'aquell individu.

## Relacio amb el tipus d'inversio

Aqu s'ha representat les mateixes dades que en els tests anteriors pero separant segons el mecanisme de generacio, per veure si el fet que siguin generades per NH o NAHR afecta en la taxa de recombinacio en la que apareix una inversio.

Sembla que en general les inversions mediades per NH apareixen en llocs de recombinació més baixa que les mediades per NAHR, que es esperable, ja que en altres articles es menciona un augment de les variants estructurals generades per NAHR en zones properes a hotspots. Tot i aixi, no sembla que hi hagi prou dades com per treure alguna cosa en clar respecte a inversions mitjanes i grans, perque nomes hi ha una inversio de tipus NH en cada grup.

## Efecte del genotip sobre la taxa de recombinació, inversió a inversió

Én la Figura 7 es pot veure el context de les inversions del grup 2 i 3.

Agafem la diferencia entre la mitjana de fora i el valor de dins. Al ser quantils em donen informacio del % de distancia entre ells. Hauria de ser una mesura comparable entre individus. En la Figura 9, s'agrupen en diferents boxplots aquestes diferències. S'ha restat el valor de dins al de fora, per tant un 0 vol dir que

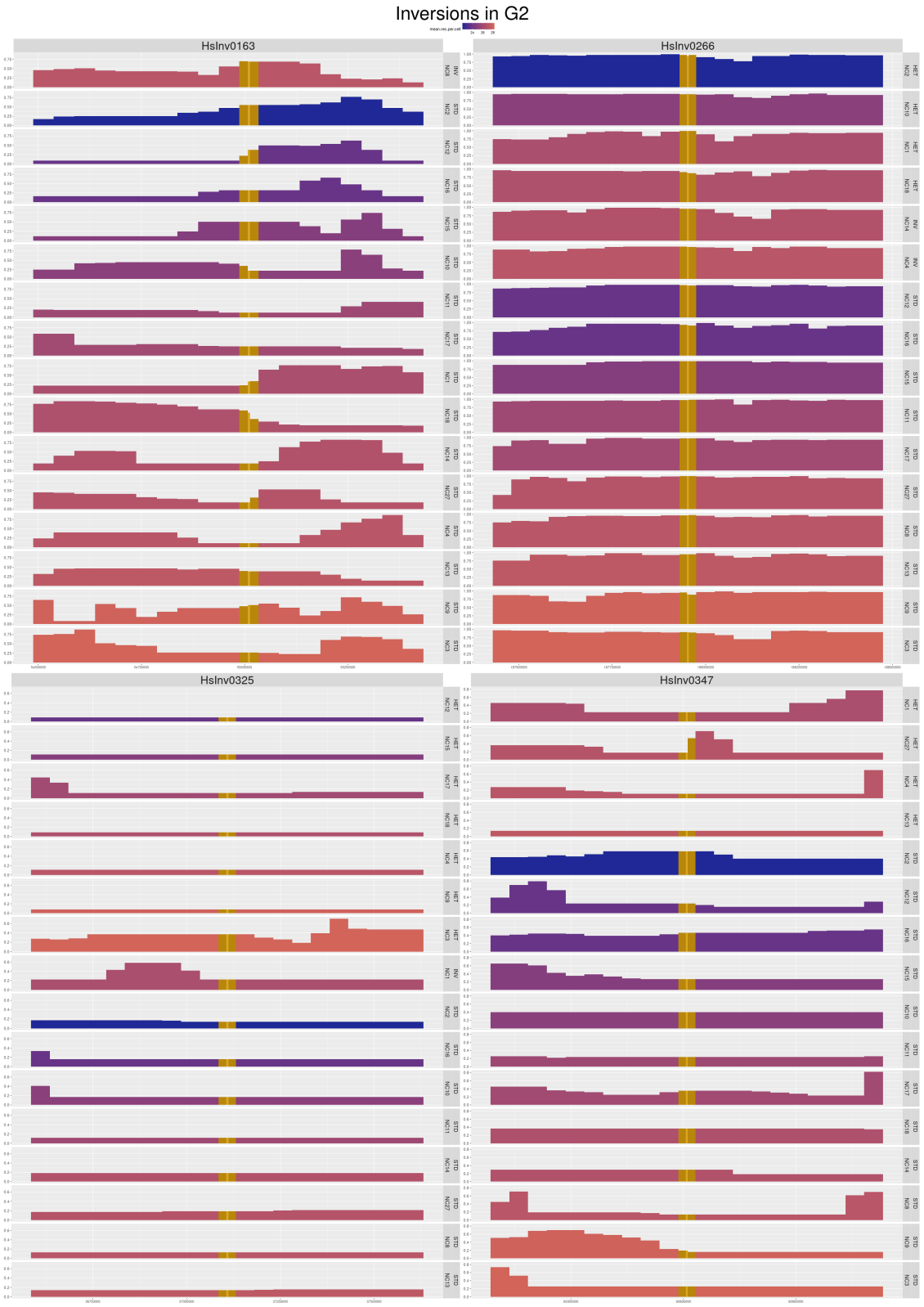


Figure 7: Inversions del grup 2 en context. Les inversions del grup 2 son d'entre 5 i 50kb. La zona interna de la inversió (finestra "in") està represenada en groc. A banda i banda, en groc fosc, estan les finestres "buffer", que inclouen els breakpoints i inverted repeats coneguts per les inversions, i per les que s'ha fixat una mida mnia de 20kb. La resta de finestres son les finestres "out". Per a cada inversio hi ha 10 finestres "out" de 50kb. Les finestes "out" estan colorejades segons la taxa de recombinacio mitjana de l'individu, i per cada inversio els individus estan endreçats per genotip i taxa de recombinació mitjana.

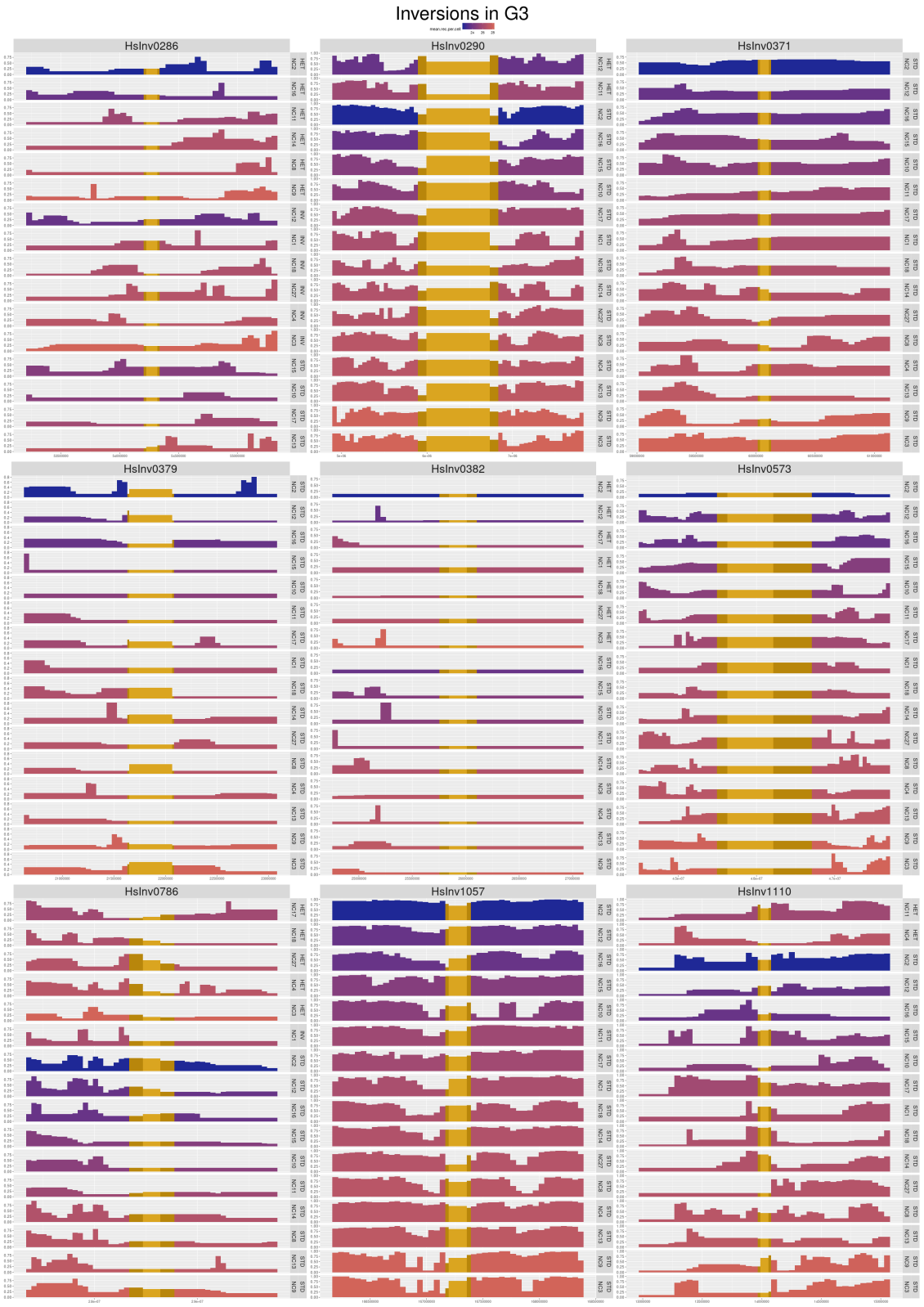


Figure 8: Inversions del grup 3 en context. Les inversions del grup 3 son de mes de 50kb. La zona interna de la inversió (finestra "in") està represenada en groc. A banda i banda, en groc fosc, estan les finestres "buffer", que inclouen els breakpoints i inverted repeats coneguts per les inversions, i per les que s'ha fixat una mida mnia de 20kb. La resta de finestres son les finestres "out". Per a cada inversio hi ha 20 finestres "out" de 50kb. Les finestes "out" estan colorejades segons la taxa de recombinacio mitjana de l'individu, i per cada inversio els individus estan endreçats per genotip i taxa de recombinació mitjana.

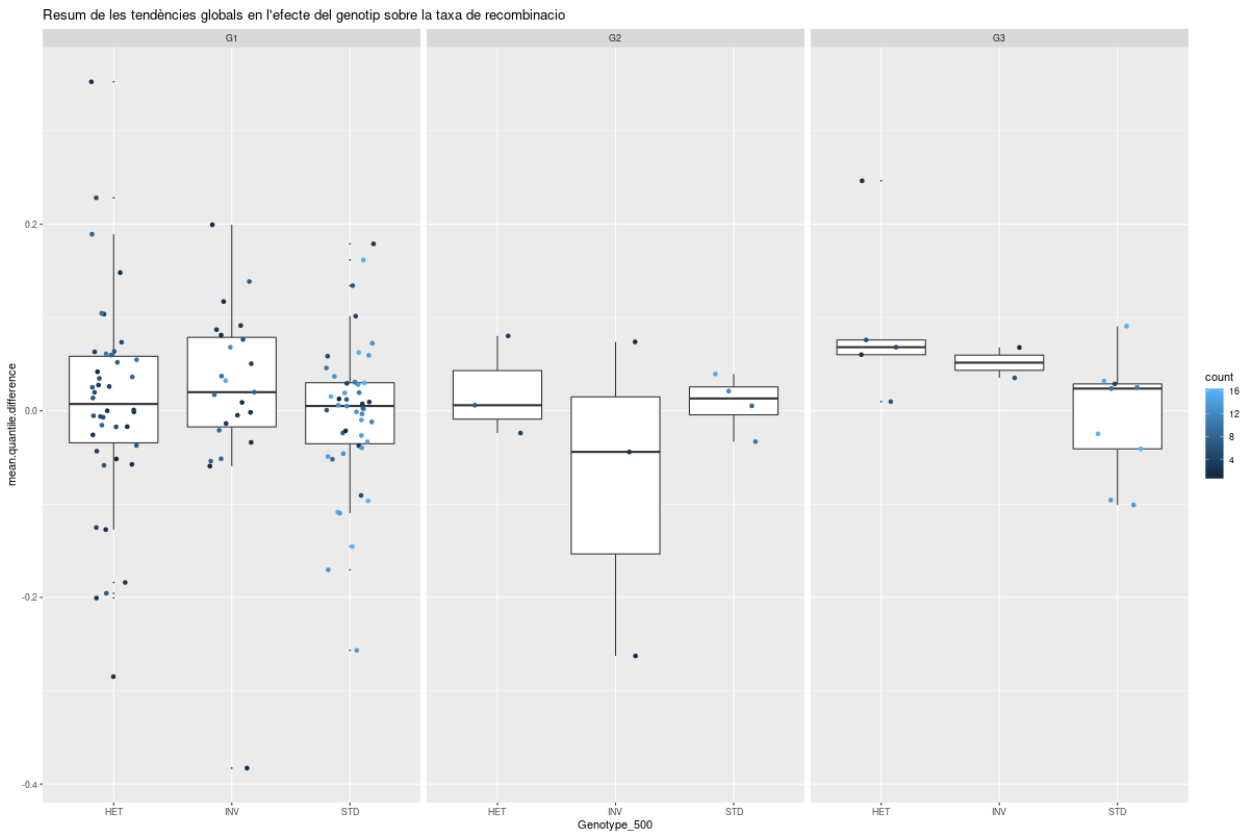


Figure 9: Efecte dels genotips de les inversions sobre la taxa de recombinacio. Cada punt es la mitjana per a una sola inversio del valor mesurat per tots els individus amb el genotip especificat. El valor mesurat es la diferencia entre la taxa de recombinacio dins la inversio i la mitjana de les finestres a fora (fora-dins). El color del punt indica en base a quants individus s'ha fet aquesta mitjana i per tant ens dona una idea sobre la fiabilitat d'aquesta dada.

ambdues mesures son iguals. Valors positius volen dir que les mesures dins de la inversio son menors que la mitjana de fora. Valors negatius, al reves, dins hi ha mes recombinacio que fora.

Com a recordatori, hi a 9 inversions del grup 3, 4 del grup 2 i 51 del grup 1, pero no totes elles tenen informacio de tots els genotips. En general, les mesures dels grups 1 i 2 mostren que no hi ha practicament diferencia entre dins i fora de la inversio. En les inversions grans, es veu que els individus estandard tenen valors lleugerament mes alts a fora que a dins, pero que aquestes diferències s'accentuen en els individus heterozigots, tot i que la mida de mostra sol ser mes grans per estandards que invertits dins de cada una de les inversions.

La majoria d'inversions del grup 1 son generades per NH, mentre que la majoria dels grups 2 i 3 son generades per NAHR, i les 3 generades per NH que hi ha entre els dos grups només aporten dades a un dels 3 genotips possibles.