Proposta de mètodes definitius

Ruth Gómez Graciani

Imputació dels genotips

Per assegurar que els genotips inferits per cadascun dels individus són fiables, es tindran en compte múltiples fonts d'informació:

- **Análisis previs d'imputabilitat:** el Jon va fer anteriorment anàlisis per avaluar si les inversions tenen tagSNPs, o en cas contrari, si són imputables amb fiabilitat mitjançant IMPUTE2, ja que de vegades pot retornar resultats que són de bona qualitat segons el programa, però en realitat estan errats.
- **Imputació amb IMPUTE2:** en el manual del IMPUTE2 el seu creador recomana no utilitzar com a referència només els individus que pertanyin a la mateixa població que la mostra, perquè IMPUTE2 ja té una passa en què busca aquells individus de la referència que més s'assemblin als que estem genotipant. Per altra banda, algunes inversions, en especial les recurrents, podrien confondre al programa a l'hora de triar individus de referència. Es van comparar diferents combinacions per origen de la mostra (tots els individus o bé la població concreta segons la procedència de la mostra, confirmada per PCA) i quantitat d'haplotips de referència (entre 100 i 500) i això va servir per destriar aquells genotips excel · lents (uniformes entre diferents metodologies i sempre de bona qualitat) i els que no eren tan precisos.
- **Densitat de SNPs i manteniment dels tag SNPs:** per tenir en compte las variacions en el *coverage* i la possible pèrdua de tag SNPs en algunes inversions que podrien disminuir la qualitat de la imputació, es calcularà la densitat de SNPs a la regió utilitzada per l'IMPUTE2 i es comptaran quants tagSNPs van ser efectivament seqüenciats, i d'aquests, quants segueixen sent tag SNPs si acceptem els genotips inferits pel programa.
- **Genotipació amb breakseq:** aquest mètode és altament fiable i en cas d'estar disponible, la resta d'anàlisis serveixen més de confirmació/reiteració del resultat o per mostrar el context de la regió en quant a la quantitat de SNPs disponibles i a com de semblant és l'haplotip de l'individu al d'altres amb la seva mateixa orientació per aquesta inversió concreta.

En general, es pretén ser estricte amb quins genotips es consideren prou bons com per ser inclosos en l'anàlisi estadístic. Idealment, serien aquells on s'hagi pogut fer el breakseq i coincideixi amb el resultat de la imputació, la qual hauria estat uniforme i de bona qualitat independentment dels paràmetres fets servir (a no ser que estés classificada com a no imputable, que llavors podria haver-hi discrepàncies) i que la densitat de SNPs sigui bona i els tagSNPs, si n'hi havia, s'hagin mantingut en alt desequilibri de lligament amb la inversió. En cas de no estar disponible la genotipació mitjançant breakseq, serien acceptats aquells genotips que mantinguin la resta de les condicions anteriors per a aquelles inversions que hagin estat classificades com a imputables pel Jon.

Ajustament de la resolució dels events d'entrecreuament

La mida mitjana dels events d'entrecreuament proporcionats per Bell et al. (2020) és de 400kb. Aquesta mida tan gran és deguda al poc *coverage* per cèl · lula que tenen. En el seu anàlisis, ells assumeixen que els events d'entrecreuament han ocorregut al centre de la regió, i fan servir finestres de 500kb. Nosaltres necessitem més resolució, com a mínim en el cas de les inversions petites. Per fer-ho, calculem la taxa de recombinació comptant la fracció dels events que solapen amb una finestra d'una mida concreta, enlloc del seu nombre en termes absoluts.

Aquesta és una forma senzilla d'augmentar la resolució, tot i que en alguns casos no és possible degut a la baixa quantitat d'events detectats, que depèn de la taxa de recombinació local, o a la mida massa gran d'aquests events a causa del baix *coverage* en aquella regió. Segons les proves que s'han anat fent, la mida mínima de finestra amb què es poden observar diferències entre finestres colindants per moltes de les regions que ens interessen és de ~10kb, però una resolució més realista que pot funcionar per la majoria de casos deu rondar entre les 100 i les 200kb, que per altra banda és una millora respecte a les 500kb de l'article original.

Sigui quina sigui la mida de finestra feta servir (que es discutirà més endavant), hem de conèixer els efectes de l'augment de la resolució sobre les estimes de recombinació, per assegurar-nos que no causem esbiaixos importants. Per fer-ho, es mirarà la correlació entre el mapa de l'article original i mapes elaborats amb finestres progressivament més petites (100kb, 50kb, 20kb).

Mostreig de les dades de recombinació

Es proposa dividir les 120 inversions en 3 grups diferents en funció de la seva mida per analitzar-les amb paràmetres més adients. Es considerarien inversions petites aquelles menors de 3kb (50 inversions, ~42%), mitjanes aquelles entre 3 i 50kb (54 inversions, ~45%) i grans aquelles majors de 50kb (15 inversions, ~13%). 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 dividirà en "in" (zona interna de la inversió, que exclou els breakpoints), "buffer" (zona immediatament colindant a la inversió, que inclou breakpoints, inverted repeats, etc. i 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 30kb per les inversions petites, 100kb per les mitjanes i 1Mb per les grans.

En cas de que no s'especifiqui el contrari, totes les mesures de taxa de recombinació es normalitzaran per controlar per l'efecte de l'individu i el cromosoma on es troba la inversió, dividint el resultat de la finestra entre la taxa de recombinació mitjana per aquell cromosoma en aquell individu concret.

Es descartaran aquelles mostres on s'observi que no hi ha canvis al llarg de tota la regió analitzada per una inversió i individu concrets a causa d'una baixa resolució de les mesures de recombinació locals.

Anàlisi estadístic

Les inversions apareixen en contexts recombinatoris concrets?

- H0: les inversions polimòrfiques poden trobar-se en qualsevol context recombinacional.
- H1: les inversions es troben en llocs de recombinació major o menor en funció de les seves característiques.
- Variable nominal: mida de la inversió; mecanisme de generació.
- Variable de mesura: increment de la taxa de recombinació mitjana en la regió al voltant de la inversió sense normalitzar respecte al seu valor de normalització (`raw mean(cM/Mb "out") / cM/Mb in host chromosome`).

Donat que en la zona "out" no esperem que el genotip de la inversió tingui efecte, es podrien fer servir tots els individus, genotipats o no, evitant la sobre representació d'alguna de les inversions.

Per comprovar si la mida de la inversió està relacionada amb el context on es troba aquesta, es podria calcular la correlació entre la mida i la variable de mesura triada. Per comprovar si el mecanisme de generació està relacionat amb el context recombinacional de la inversió, es pot fer un ANOVA on els grups son diferents mecanismes.

Efecte del genotip sobre la taxa de recombinació

- H0: el genotip de la inversió no té efecte sobre la diferència en taxa de recombinació dins de la inversió ("in") vs. fora ("out").
- H1: els canvis d'orientació van acompanyats de canvis en la taxa de recombinació local dins de la inversió.
- Variable nominal: genotip de la inversió (3 grups).
- Variable de mesura: increment en la taxa de recombinació dins de la inversió respecte fora (`cM/Mb "in" / mean(cM/Mb "out")`).

En principi aquesta hipòtesi es podria testejar amb un ANOVA o alguna de les seves variats. El principal problema que em trobo a l'hora de fer aquest anàlisi és que no tinc la mateixa quantitat de individus genotipats ni genotips de cada tipus per cada inversió. En cas de fer un one-way ANOVA tenint en compte els 3 possibles genotips dels individus, no és obligatori tenir grups de la mateixa mida a no ser que les desviacions estàndard entre diferents genotips siguin molt similars, pero no sé com afectaria la sobre representació d'algunes de les inversions (ex. la inversió A té 3 individus analitazts per cada genotip i la inversió B té 1 sol individu analitzat). Un two-way ANOVA on afegissim la inversió com a variable nominal no soluciona el problema, ja que totes les combinacions entre els diferents grups han d'existir (per tant descartem aquelles que només tenen mostres per 1 o 2 genotips) i a més, tot i que existeix la possibilitat de tenir grups de diferents mides, això dificulta molt l'anàlisi. Finalment, un nested ANOVA sembla que no és tan sensible canvis en la mida dels grups, pero els subgrups han de ser únics del grup principal, és a dir, que no podria mirar la mateixa inversió per diferents genotips. En general, penso que considerar dividir el genotip en 2 grups (STD vs. HET per la inhibició de la recombinació de l'heterozigot i STD vs. INV per canvis en els patrons en la taxa de recombinació causats pel canvi d'orientació) podria resoldre part dels problemes, pero no tots.

Altres anàlisis

Podria ser interessant repetir l'anàlisi de l'efecte del genotip sobre la taxa de recombinació pero comparant "in" amb "buffer" per veure l'efecte del genotip sobre el seu entorn immediat. Potser s'haurien d'analitzar per separat les inversions segons el seu tipus de "buffer" (breakpoint petit, breakpoint gran, amb inverted repeats...)

També es poden fer figures o anàlisis més exhaustius per les inversions més grans, les quals es poden dividir en finestres 5 finestres proporcionals a la mida de la inversió, amb el "out" dividit en finestres de 100kb, ja que es podria esperar un efecte més clar i de més llarg abast en les inversinos de mida gran.