Mario Caceres dt., 31 de març 12:22

per a Ruth, Jon

Hola Ruth,

Espero que vayan bien estos días y disculpa el retraso en contestar, que he estado liado con proyectos de coronavirus. Supongo que ya habrás comentado los resultados con Jon y lo pongo en copia como experto en los temas de imputación.

A mi me ha parecido muy bien el report y he añadido algunos comentarios. En principio yo creo que se trata de acabar de decidir que inversiones consideramos como que se imputan bien y a partir de aquí se puede seguir adelante con los análisis. Si que es verdad que para muchas inversiones hay pocos individuos heterocigotos, pero ya es lo esperado dado que partiamos de 20 muestras y tendriamos que pensar como se pueden hacer los análisis combinado inversiones (todas juntas o por ejemplo según el tamaño y mecanismo de origen para tener más poder).

Las principales cosas a comentar serían:

- No se si se podría estimar mejor el error en la imputación en la diferentes estrategias con los datos de los individuos con genotipo conocido o al menos comparar mejor con los datos de Jon. Entiendo que eso debe ser la columna de "known imputability" del excel, pero no estará mal tener los datos del % de error. El problema es que no se como se puede estimar en los individuos admixed cual de las 3 estrategias va mejor, porque diría que nosotros no tenemos ninguno genotipado.

- Para las que tienen tag SNPs no se si has comprobado que los genotipos imputados coinciden con los imputados.

- En la 58, como es que no está el tag SNP? No se puede recuperar de alguna manera para ahorrarnos el breakseq? De hecho si se hace el breakseq ya se puede hacer para todas las de breakpoints sencillos, aunque supongo que la mayoría se imputan bien.

Saludos a los dos,

Mario

Jon Lerga Jaso dt., 31 de març 17:35

per a Mario, Ruth

Hola a los dos,
Muy chulo el informe! Qué profesionalidad. Lo he revisado junto con la tabla de genotipos imputados que me pasaste el otro día, Ruth, pero que todavía no había mirado.
Cosas a comentar:

* Breakseq y tagSNPs
- Por un lado, como dice Mario, estaría bien revisar que todos los genotipos imputados coinciden con los tagSNPs. Se puede correr en un momento PLINK con los VCFs con genotipos imputados por población y mirar directamente si los tagSNPs siguen siéndolo. Por si te sirve:

```
# Linkage disequilibrium between HsInv0124 and other variants
folder_plink="/home/jon/soft/
cd $folder_files/Linkage_
vcftools --gzvcf $folder_files/VCF/Genotypes_
$folder_plink/plink --file genotypes_plink_format --r2 --ld-snp HsInv0124 --ld-window-kb 1000000
--ld-window 999999 --ld-window-r2 0 --noweb
```

- Por otro, el tema del Breakseq.
Si no recuerdo mal, no estaban los tagSNPs de la INV58 en tus VCFs. De hecho, en la tabla de imputación ya se ve que no hay manera con la 58... Y creo que comentamos que alguna otra INV con tagSNPs tampoco si imputaba perfectamente, no?
Se puede usar el breakseq, que tampoco es muy dificil de poner. Yo esta semana estoy liado, pero si eso, la semana que viene podemos mirarlo, y ponerlo. Y ya que estamos para la 58, lo ponemos para las ~20 inversiones que tienen librerías.

Así, quedará bastante completo: breakseq + imputation + tagsnps
Y vemos si coinciden los genotipos de todas las estrategias entre sí.

* Acabar de decidir los genotipos:
Por un lado, se pueden dar por buenas todas las inversiones que tengan genotipos del breakseq e imputados que coinciden con sus tagSNPs.
Esto debería ser ~2/3 de las inversiones autosómicas.

Del resto, hay que hacer varias consideraciones:
- Hay inversiones como la HsInv0069 que no son imputables, aunque las estimas según IMPUTE2 sean buenas o las distintas estrategias arrojen el mismo genotipo de manera coherente. En tu excel, por ejemplo, esta inversión aparece con un genotipo con buena calidad en los individuos nc1, nc16, NC26ab; pero no es creíble.

Cuando hice el análisis de excluir una muestra con genotipo conocido y luego imputarla con el objetivo de ver cómo de bien se imputa cada inversión, para la INV69 ocurría que hay muestras que IMPUTE2 las imputa con un 100% de fiabilidad pero es un genotipo erróneo. Ejemplos:

Individuo - genotipo real - genotipo imputado - Error
NA20818 - 0/1 - 0,0,1 (= 1/1) - Het -> Inv
NA20515 - 0/0 - 0,1,0 (= 0/1) - Std -> Het

Por eso, yo **no** me fiaría de estas inversiones, aunque los porcentajes de probabilidad sean buenos y los genotipos coherentes con diferentes estrategias.

- Las que son imputables según los datos de mi tesis, aunque tengan una probabilidad moderada (~0.7) según IMPUTE2, si el genotipo es coherente, yo lo daría por bueno.
Ejemplo: la 266.

Yo quizá haría lo siguiente:

(1) Comprobar cómo se imputan las inversiones en EAS/SAS mediante la estrategia de excluir una muestra con genotipo conocido e imputarla. Si no recuerdo mal, te pasé el script que yo usé para mi tesis. Puedes adaptarlo para la población asiática, aunque hay que tener en cuenta que este paso llevará tiempo! No necesitarías hacer las inversiones que tengan tagSNPs, solo las que hay que ver si se imputan bien o no.

(2) Aquellas inversiones que no se imputen bien en ninguna población (EUR, AFR, EAS/SAS) según las pruebas de imputación, excluirla.
Si no se imputa bien en alguna población, excluir la inversión de los individuos con mezcla de poblaciones.
Probablemente, por ejemplo, las inversiones 228 o la 69 se excluirían (no se imputan en AFR ni en EUR), entre otras.

(3) Las que se imputen bien en distintas poblaciones y tengan genotipos con una probabilidad aceptable (>0.7?), nos las quedamos.
Si la probabilidad es baja, aunque sea imputable, excluir el genotipo también.
En este caso, nos quedaríamos inversiones que se imputan bien en algunas poblaciones, pero no nos quedaríamos el genotipo de las poblaciones que no se imputarían bien (por ejemplo, para la 30, en EUR sí, en AFR no).

<u>* Tabla 2 - INVs & DELs:</u>

No veo necesidad de hacer nada especial en estas inversiones. Por ejemplo, la INV 52 tiene tagSNPs y la DEL también. Serían dos entradas distintas en el VCF:

- chr pos INV52 Std Inv ..... 0|0 0|1 1|1 ...
- chr pos DELX Ref Alt ..... 0|0 1|0 0|0 ...

De ahí ya deduces donde la inversión ha sido delecionada. Por tanto, seguimos con el mismo criterio: (1) STD/INV para la INV y (2) REF/ALT para la DEL.

* Genotyping PCA:

Veo que la PCA la has hecho solo con el chr1. En principio, es suficiente resolución, pero lo ideal sería con todos los autosomas, de cara a un futuro paper, para separar completamente las muestras.
No sé si has mirado la componente 3, además de la 1 y 2; quizá ayude para clasificar mejor, aunque parece que las muestras ya están bien clasificadas.
Supongo que las muestras a las que se les ha asignado una población caen claramente en el cluster correspondiente. Quizá, se puede usar algún programa como ADMIXTURE (Alexander et al.,2009) o EIGENSTRAT (Patterson et al., 2006), pero no lo veo necesario.

* Otros comentarios menores sobre el informe:

- ...500, 250 or 100 closest individuals to those in test panel (the second methodology with **relaxed, medium and strict conditions**...
Yo no llamaría a estas condiciones así. Simplemente k=100 haplotypes, k=250...

- En la PCA se pueden mostrar las 20 muestras, quizá, para que se vean bien donde caen.

- El reference panel en hg19, no **hg38**

- "**qctools** pca" -> "qtltools pca" ?
- "**icludes**" -> "includes"

Ánimos, que ya estamos en el ecuador del confinamiento (si no lo alargan, claro).
Jon

Mario Caceres dt., 31 de març 18:40

per a Jon, Ruth

Muchas gracias Jon!

Por mi parte el plan me parece muy sensato! Sólo un par de puntualizaciones:

- La estima del error de la imputación en EAS/SAS quizás estaría bien hacerla exactamente igual que la hiciste tu para las otras poblaciones en tu tesis. Lo digo por si lo podiamos aprovechar para el paper funcional. En principio no hacía falta porque en asiáticos no había datos funcionales, pero yo creo que estaría bien dar una idea de cuan bien se imputan las inversiones en las diferentes poblaciones y cuando lo leí lo eché en falta. Una pregunta, si al final se añade alguna inversión más al estudio, hay que rehacerlo todo? O se puede simplemente repetir para las inversiones nuevas?

- Respecto al brekaseq, un problemilla es que creo que había que alargar la secuencia de las librerías y no lo acabé. De todas formas de momento nos podemos centrar en las inversiones que den problemas de imputación, que no serán muchas y así en un momento está hecho. Como son sólo 20 individuos no es un gran problema si luego hay que repetirlo con más inversiones.

Ánimos a los dos que ya queda menos!

Mario

Ruth Gómez

per a Jon, Mario

Hola Mario y Jon,
Gracias a los dos por las sugerencias! Contesto a los comentarios:

- **La PCA:** En la Tabla 1, cuando en la columna 'PCA result' pone una población concreta, ese individuo hasta donde yo he podido ver pertenece a esa población. En la Figura 1 puse los que podrían ser más dudosos, que son aquellos admixed y en los que cambia la población respecto a la información que daba la clínica de fertilidad (excepto el 27, que era SAS de forma muy clara). La figura final sería con todos. Miré todas las combinaciones entre las componentes 1, 2 y 3. Como en principio estos individuos no los cambiaremos, puedo hacer ya la PCA con todos los autosomas y así se queda hecho.
- **Inversiones con tag SNPs:** Comprobé manualmente si la imputación y el genotipo de los tag SNPs coincidía para algunas de las inversiones y he encontrado ejemplos en los que algunos supuestos tag SNPs con un genotipo que no concordaba con el resto y por tanto con el genotipo imputado, así que haré como sugiere Jon y miraré si los tagSNPS siguen siéndolo.
  En general, las inversiones que tienen tag SNPs o que son imputables han salido bien y las que no son imputables han salido mal. La 58 es la única con tag SNPs que ha salido realmente mal, pero hay otras que no han salido tan bien como el resto (la 156 y la 991). Como menciona Jon, en las que se supone que no son imputables y deberían salir mal también hay excepciones, como la 796 y la 371, pero como digo más adelante la idea con estas era descartarlas, aunque las haya imputado igualmente, un poco para ver si las que tenían que salir mal salían mal.
- **Breakseq:** Sobre la 58, al principio pensaba que sería problema de la calidad de los SNPs, ya que con IMPUTE2 sólo he usado aquellos que tenían el tag 'PASS' en el VCF. Sin embargo, al comprobar manualmente el VCF, en él no aparece ninguno de los tag SNPs para la 58, ni de buena ni de mala calidad. Coincido con que si se hace el breakseq se aplicaría a todas las que se pueda, que será más fiable que la imputación servirá además de control de calidad.
- **Decidir los genotipos:**
- *No imputables:* Ya me comentaste que las no imputables aunque saliesen bien no nos podíamos fiar, y aunque las he analizado y tenido todas en cuenta aquí, tenía planeado filtrarlas.
- *Imputables y con tag SNPs:* La idea de usar las 4 estrategias para todos los individuos era precisamente comparar, en aquellos que sabemos su población, si usar de referencia todos los individuos o poner solo los de la población concreta funciona igual de bien, de cara a interpretar los individuos admixed. En general, hay pocos casos en los que inversiones imputables o con tag SNPs difieran en los resultados de uno a otro método. Respecto a si es mejor coger k = 500, 250 o 100 individuos, no parece que una k vaya mejor que la otra. Normalmente o todas funcionan o todas fallan, y si sólo una funciona no es siempre la misma. Por tanto, como dice Jon, iría caso por caso:
- En los individuos en los que se sabe la población, aunque haya discrepancias respecto al panel con todos los individuos, me quedaría con lo que dice su propia población si tiene una probabilidad >70-80%.
- En en los individuos admixed y el SAS (cuando falla por falta de referencias), me quedaría con aquellos que: 1) se imputan bien siempre en los otros individuos y 2) han salido bien independientemente de la k que hayamos usado.
  Me parece bien comprobar cómo se imputan las poblaciones asiáticas, para poderlas clasificar como Imputables y No imputables a la hora de decidir los genotipos porque si no, es verdad que eso se

queda un poco en el aire. Además uno de los individuos admixed tiene parte de asiático, o sea que en ese caso ni siquiera me habría podido fiar del todo usando aquellas que sean imputables en EUR y AFR.

- **INVs y DELs:** No sé si me he explicado mal en el report, pero precisamente he hecho eso que dice Jon.
- **Inversiones nuevas:** Con los datos necesarios, en cualquier momento se pueden hacer inversiones nuevas. De hecho, hice primero las 45 “de siempre” y después Jon me pasó las nuevas. Al final analizé los dos datasets juntos por no tener que ir moviéndome entre varios archivos pero el código está preparado para analizar inversiones nuevas si vamos teniendo más datos.
  Como ha dicho Jon, la semana que viene nos íbamos a mirar lo del breakseq. Mientras tanto yo estaba planteando el análisis estadístico para poder combinar las inversiones y mirándome lo de las networks de Carla. Como resumen, a parte de eso tendría que re-hacer la PCA, mirar si los tag SNPs lo siguen siendo y estimar el error de la imputación en individuos asiáticos.

  Espero no dejarme nada!

  Saludos,

  Ruth

Mario Caceres dc., 1 d’abr. 10:57

per a Ruth, Jon

Buenos días Ruth,

Muy bien por el plan! La 58 puede ser que como está en una zona un poco complicada hayan filtrado todos los SNPs. Si quereis cuando esté claro que inversiones sin inverted repeats (no NAHR) dan problemas, me las enviais y acabo de revisar las librerías para hacer el breakseq.

Ánimos!

Jon Lerga Jaso dc., 1 d’abr. 11:56

per a Mario, Ruth

Yo alargué las librerías de algunas inversiones cuando estuve en estonia. Tenemos 9 con 300 bp (antes 100bp). Entre ellas está la 58.
Ruth, los reads de tu proyecto eran de 191 bp no? De momento, podemos tirar con estas, que son de un tamaño adecuado.
La semana que viene lo miramos.

Ruth Gómez Graciani <ruth.gomez.graciani@uab.cat> dc., 1 d’abr. 14:43

per a Jon, Mario

De acuerdo! Si, eran de 192 bp.

Mario Caceres dc., 1 d’abr. 17:03

per a Ruth, Jon

Vale, y si hay que alargar alguna me avisais, que tengo todas las secuencias en el ordenador y lo puedo hacer rápido.