20 de abril de 2016
Consuelo Dayzu Quinto

Concordancia entre genotipos de MEGA (Illumina) y 1KG (Affymetrix)/HapMap (Affymetrix+Illumina)

A continuacion se muestran los resultados de 3 rondas de analisis de la corrida de entrenamiento del HiScan, con el microarreglo MEGA de Illumina.

Cuadro 1: Muestras usadas en la corrida de entrenamiento

ID muestra	Plataforma
NA21402	HapMap
NA12878	HapMap/1KG
NA21405	HapMap
HG01938	1KG
NA21685	HapMap
HG01941	1KG
NA19088	HapMap/1KG

Cada ronda de analisis consiste en los siguientes pasos:

- Los genotipos fueron generados usando GenomeStudio con el manifiesto MEGA_Consortium_15063755_B2, y con las siguientes variaciones:
  1. sin cluster file
  2. PAGE cluster file (`https://www.pagestudy.org/index.php/multi-ethnic-genotyping-array`)
  3. GLOBAL cluster file (en la computadora de HiScan: MEGA-Global:multi-ethnic-global-8-cluster-file/Multi-EthnicGlobal_ClusterFile.egt)
- SNPs en cromosomas 0, X, Y, XY y MT fueron filtrados, al igual que SNPs con missing call superiores a 0:
  `plink –file [archivo] –geno 0 –not-chr 0,X,Y,XY,MT –recode –out [archivo]`
- Los identificadores de los SNPs fueron renombrados a rsID usando el archivo MEGA_Consortium_v2_15070954_A1_b138_rsids.txt.
- SNPs con el mismo rsID y la misma posicion fisica tambien fueron removidos:
  `plink –file [archivo] –exclude [lista duplicados] –recode –out [archivo]`
  `plink –file [archivo] –list-duplicate-vars`
- Para cada cluster file usado, se hizo una comparacion con los genotipos obtenidos por 1000 Genomes (1KG) y HapMap:
  1. 1KG: Affymetrix 6.0 (863, 597 SNPs)
  2. HapMap3: Illumina Human1M y Affymetrix SNP 6.0 (1, 374, 871 SNPs)
- Se encontraron los SNPs en comun entre cada una de las plataformas (busqueda por rsID) y se omitieron SNPs que tuvieran por nombre ".".
- Con plink, se unieron los diferentes sets de datos (entrenamiento+1KG y entrenamiento+HapMap) y se cambiaron las cadenas (AC a TG) para algunas variantes:
  `plink –bfile [archivo1] –bmerge [archivo2.bed/bim/fam] –recode –out [archivo3]`
  `plink –bfile [archivo2] –flip [archivo2.missnp] –make-bed -out [archivo2]`
  `plink –bfile [archivo1] –bmerge [archivo2.bed/bim/fam] –recode –out [archivo3]`

- Para comparar los genotipos de los mismos individuos en las diferentes plataformas, se uso el script cal_concordance.pl, el cual compara los genotipos y calcula el numero de diferencias:
  $matches = ($split1[1] ^$split2[1]) =~ tr/\0//;

En la siguiente tabla estan los resultados de concordancia entre los diferentes analisis:

	Concordancia con HapMap			Concordancia con 1KG		
	No cluster	cluster PAGE	cluster GLOBAL	No cluster	cluster PAGE	cluster GLOBAL
% concordancia (promedio)	96 %	94 %	94 %	94 %	92 %	92 %
# SNPs	282, 831	279, 879	268, 907	128, 587	127, 318	122, 581

En las siguientes figuras se muestran los porcentajes de concordancia (en naranja) entre los genotipos de los individuos en la corrida de entrenamiento, en 1KG y en HapMap.

Figura 1: Comparacion sin cluster

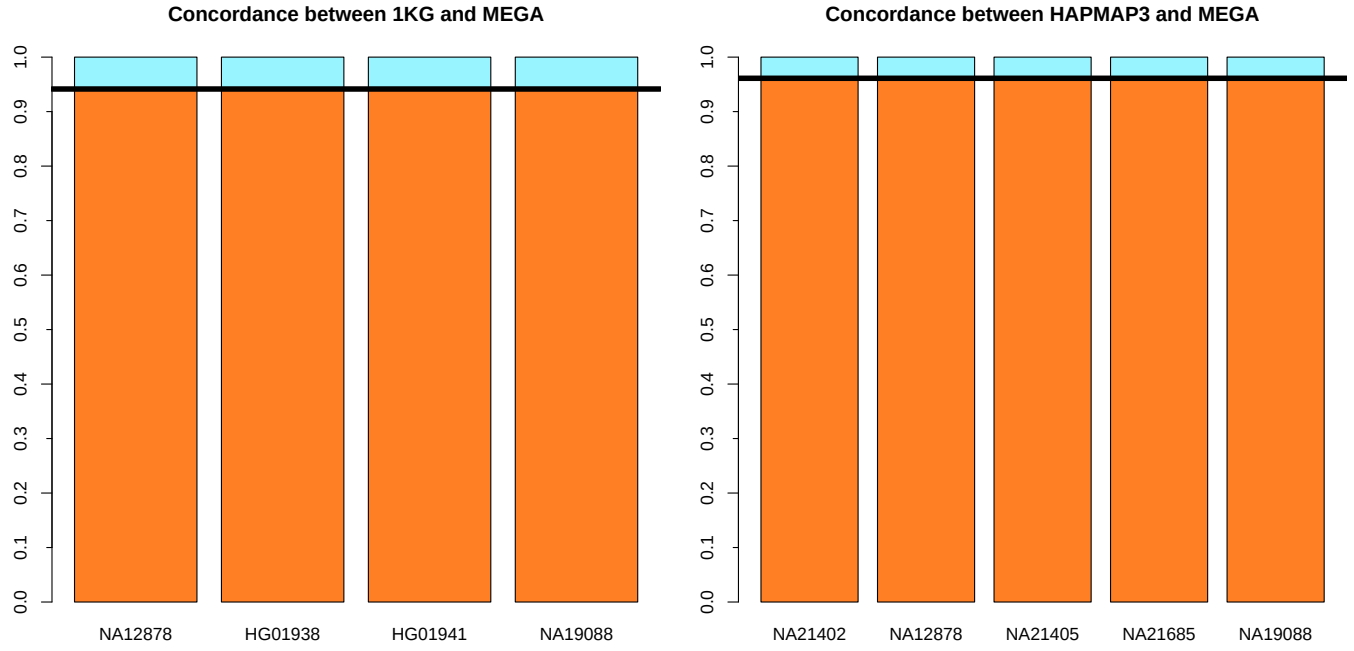


Figura 2: Comparacion cluster PAGE

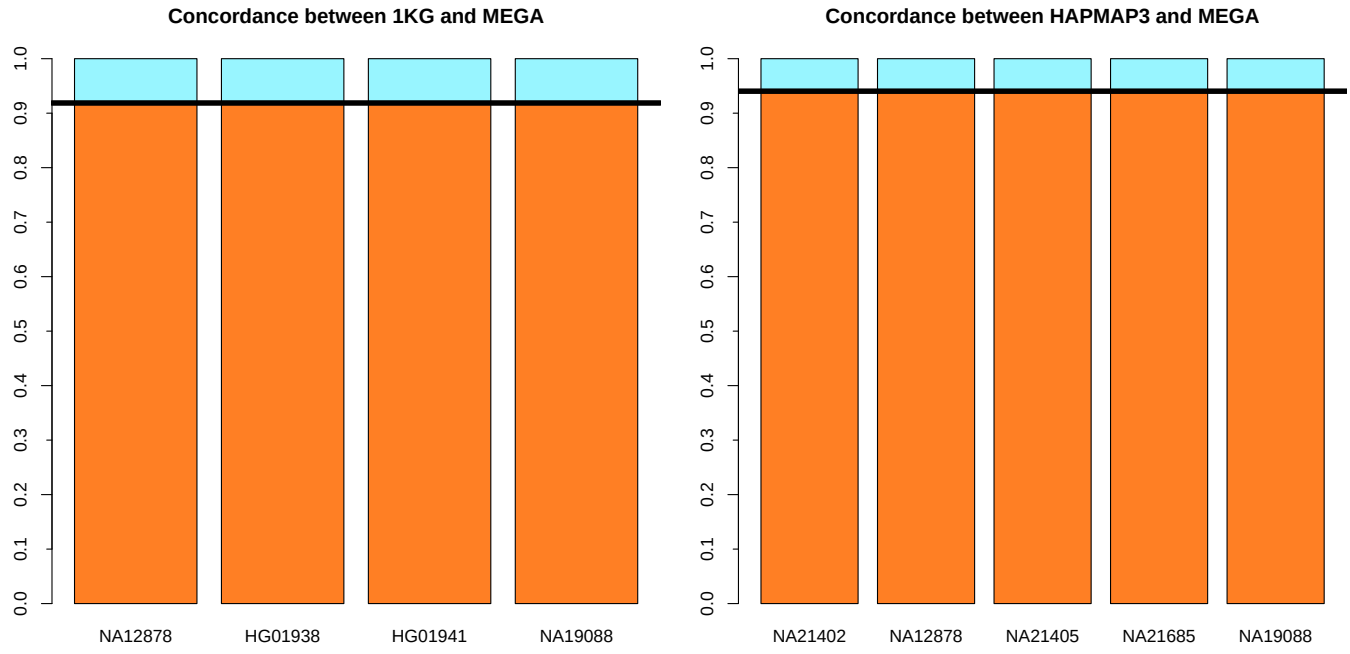
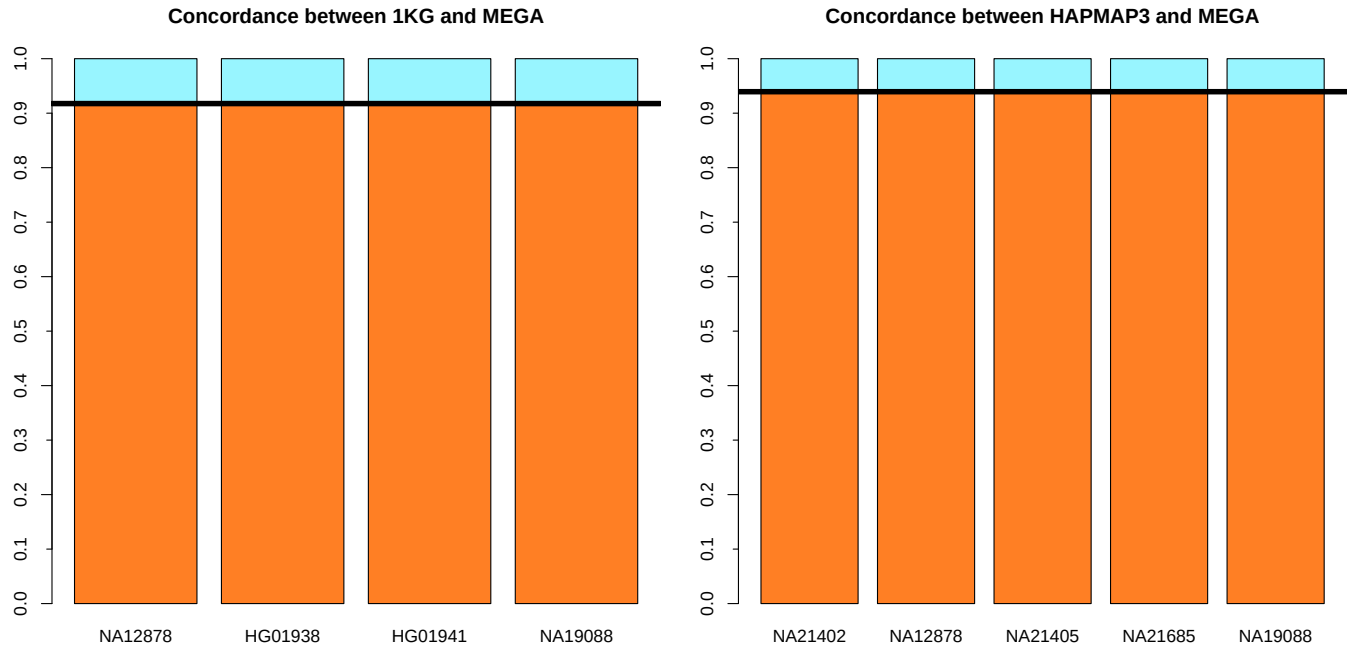


Figura 3: Comparacion cluster GLOBAL



(Los archivos README tienen mas informacion acerca de los pasos que se siguieron para renombrar los SNPs, filtrar y unir los datos.)*

En base al analisis realizado en el primer batch de muestras de Maria Avila y 1KG, me di cuenta que hay sitios que tienen, por ejemplo, un genotipo CC, mientras que en el segundo set de datos, el mismo genotipo es GG (lo mismo aplica para genotipos AA y TT). Con el script cal_concordance2.pl, se tomaron en cuenta este tipo de sitios y los nuevos porcentajes de concordancia estan el siguiente tabla:

	Concordancia con HapMap			Concordancia con 1KG		
	No cluster	cluster PAGE	cluster GLOBAL	No cluster	cluster PAGE	cluster GLOBAL
% concordancia (promedio)	99.81 %	99.96 %	99.96 %	99.82 %	99.97 %	99.98 %
# SNPs	282, 831	279, 879	268, 907	128, 587	127, 318	122, 581

GenomeStudio e informacion acerca de la posicion de los SNPs (cadena positiva/negativa)

En las corridas de analisis de Genome Studio con las muestras del entrenamiento y usando el manifiesto MEGA_Consortium_15063755_B2.bpm, no se muestra en la SNP Table la informacion acerca de la cadena donde se encuentran los SNPs (+ vs -).

Sin embargo, esta informacion si aparece cuando se analizaron las muestras de Maria Avila con el microarreglo de MEGA comercial y el manifiesto Multi-EthnicGlobal_A1.bpm. A pesar de esto, parece que la informacion de la direccion de las cadenas no se transmite a plink, ya que los archivos creados por el plugin de plink siguen teniendo alelos en la cadena - (TT vs AA).

Ejemplo:
rsId: rs2465136
posicion: 1: 990417
muestra: HG01938 (1KG)

Informacion de dbSNP: A/G (REV)
Alelos en 1KG: TT
Alelos en MEGA reportados por GenomeStudio: AA

Algo que se puede hacer para encontrar sitios donde esto ocurre es generar un reporte final desde GenomeStudio donde se encuentra la informacion acerca de la direccion de las cadenas (Plus/Minus Strand), pero solo usando el manifiesto Multi-EthnicGlobal_A1.bpm.

Una sugerencia de Pavel es generar una lista de SNPs (basada en el reporte final) donde potencialmente habria un problema de cadena + vs. cadena -, para que el usuario sepa que esos alelos se tienen que cambiar con plink (opcion –flip lista.txt) cuando se querian unir varios sets de datos como 1KG y/o HapMap.