GWAS dataに対するwhole genome imputation実施方針 （ver1.1）

2011/12/10　理化学研究所　統計解析研究チーム　岡田随象

①：概要
・複数施設で実施されたGWASを対象としたメタアナリシスにおける、whole genome imputation実施手順を記載する。

②：基となるGWASデータ
・各施設でタイピングされたGWASにおけるジェノタイプデータを対象とする。
・ジェノタイプデータに対するQuality　Controlは、各GWASにおいて採用された基準を適用する。

③：Referenceデータ
・Phase II HapMap JPT+CHB data (release 24, non-redundant, forward strand)を採用する。
・tri-allelic SNP, JPT+CHBにおけるmonomorphic SNPを除外する。（約250万SNPがimputation対象になる。）

④：Imputation手順
・使用ソフトはMACH version 1.0 （ http://www.sph.umich.edu/csg/abecasis/MACH/index.html ）とする。
・結果の再現性を担保するため、各施設内においては固定したRecombination/error rateをパラメータとして用いる。

④-1：Reference haplotypeのPhasing
・QC済のreference　HapMap dataを、Contig間のギャップ（NCBI Build 36.3）に基づき分割した後（**Table 1**）、Phasingする。
・各Contigを実行単位とすると、対象領域の長さに差が生じるため、「Contig間のギャップに基づく実行単位を採用する」という方針とする。
・引数は “-d hoge.dat –p hoge.ped --rounds 50 --states 200 --phase”とする。

④-2：GWASデータの整備
・QC済のGWASデータから、QC済のreference dataに含まれないSNPを除外する。（これらのSNPは再利用しない。）*
・GWASデータのアレルと、HapMapデータのアレルを一致させる**
・ReferenceデータとGWASデータとで、SNPの位置情報が異なる場合、referenceデータを優先する。

*：少数であること（ex. <10,000SNP for illumina 610k）、異なるplatform間でのメタアナリシスにおいては最終的に除外される可能性が高いことを考慮。
**：アレル1/2の表記は各施設の方針を採用し、メタアナリシス実行者がすり合わせる方針とする。

④-3：Recombination/error rateの推定
・各施設のおいて、（A）または（B）のどちらかを採用する。
・（A）：④-1： Reference haplotypeのPhasing時に得られた”hoge.rec”及び”hoge.erate”ファイルを採用する。
・（B）：GWASデータから一部のサンプルを抽出し、各実行単位毎に再度Recombination/error rateを推定する。
・（B）における引数は “-d hoge.dat –p hoge.ped –h hoge.haplos –s hoge.snps --rounds 50 --greedy --geno”とする。

④-4：Imputaionの実施
・全GWASデータに対して、各領域単位毎にimputationを行う。
・引数は “-d hoge.dat –p hoge.ped –h hoge.haplos –s hoge.snps --crossovermap hoge.rec --errormap hoge.erate --greedy --mle --mledetails”とする。

⑤：Imputation結果のまとめ方
・⑤-1：SNP別の各種統計量ファイル、⑤-2：”Best-guess genotype”に基づくジェノタイプファイル、⑤-3：”expected number of allele dosage”に基づくジェノタイプファイル、の三種類のファイルで結果をまとめる。

⑤-1：SNP別の各種統計量ファイル
・各SNP毎に、下記の項目を記載する。
rsID、染色体番号、染色体上の位置(bp)、アレル1/2に対応する塩基*、imputation後MAF**、Quality**、*Rsq***、GWASデータ上に含まれる（=1）か否か（=0）。

*：各施設におけるアレル1/2の表記、及びforward strandに基づく
**：Imputation実施時に得られたhoge.mlinfoファイル上の情報を記載する。

⑤-2：”Best-guess genotype”に基づくジェノタイプファイル

・Imputation実施時に得られたhoge.mlgenoファイル中の”Best-guess genotype”に基づき、ジェノタイプファイルを作成する。
・ hoge.mlgenoファイル中のジェノタイプを、”1/1”=0, “1/2” or “2/1”=1, “2/2”=2、で表記する。
・ hoge.mlgenoファイル中のアレル1/2はreference HapMap上のアレル1/2に対応するため、即ち、アレル2の本数（整数）を表記することになる。
・縦列がサンプル、横行が各SNPに対応した、タブ区切りテキストファイル形式で保存する。*

⑤-3：”expected number of allele dosage”に基づくジェノタイプ

・推定されたジェノタイプ別事後確率による”expected number of allele dosage”に基づくジェノタイプファイルを作成する。
・ hoge.mldoseファイル中のallele dosageに基づき、0-2の間に分布するアレル2の期待本数（小数）で表記する。
・ hoge.mldoseファイル中のallele dosageは、hoge.mlinfoファイル中のA1アレルの期待本数を表す。
・⑤-2のジェノタイプファイルのと整合性をとるために、下記の変換を行う。

hoge.mlinfoファイル中のA1アレル==“1”のSNP ⇒ 2– allele dosage.
hoge.mlinfoファイル中のA1アレル==“2”のSNP ⇒ allele dosage.

・縦列がサンプル、横行が各SNPに対応した、タブ区切りテキストファイル形式で保存する。*

⑥：本資料に基づくwhole-genome Imputation結果を採用したスタディの一覧

· **Okada Y**, Sim X, Go MJ, Wu JY, Gu Det al. (2012) Meta-analysis identifies multiple loci associated with kidney function-related traits in east Asian populations. ***Nat Genet.*** 44:904-909.
· **Okada Y**, Terao C, Ikari K, Kochi Y, Ohmura K et al. (2012) Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. ***Nat Genet.*** 44:511-516.
· **Okada Y**, Kubo M, Ohmiya H, et al. (2012) Common variants at CDKAL1 and KLF9 are associated with body mass index in East Asian populations. ***Nat. Genet.*** 44:302-306..
· Wen W, Cho YS, Zheng W, ..., **Okada Y** et al. (2012) Meta-analysis of genome-wide association studies in East Asians identifies novel genetic variants associated with body mass index. ***Nat. Genet.*** 44:307-311.
· Gieger C, Radhakrishnan A, Cvejic A, ..., **Okada Y** et al. (2011) New gene functions in megakaryopoiesis and platelet formation. ***Nature*** 480: 201-208.
· **Okada Y**, Hitota T, Kamatani Y, et al. (2011) Identification of nine novel loci associated with white blood cell subtypes in a Japanese population. ***PLoS. Genet.*** 7:e1002067.
· **Okada Y**, Takahashi A, Ohmiya H, et al. (2011) Genome-wide association study for C-reactive protein levels identified pleiotropic associations in the *IL6* locus. ***Hum. Mol. Genet.*** 20:1224-1231.

ご不明な点や、ご質問は **yokada-tky (at) umin.ac.jp** までご連絡下さい。

Table 1

	start	end	length		start	end	length
1pA	1	29,750,669	29,750,668	9p	1	47,107,499	47,107,498
1pB	29,800,670	121,186,957	91,386,287	9q	65,207,500	140,273,252	75,065,752
1qA	141,476,958	204,189,330	62,712,372	10p	1	39,194,941	39,194,940
1qB	204,239,331	247,249,719	43,010,388	10q	41,674,942	135,374,737	93,699,795
2p	1	91,689,898	91,689,897	11p	1	51,450,781	51,450,780
2qA	94,689,899	149,398,828	54,708,929	11q	54,450,782	134,452,384	80,001,602
2qB	149,498,829	242,951,149	93,452,320	12p	1	34,747,961	34,747,960
3p	1	90,587,544	90,587,543	12q	36,142,962	132,349,534	96,206,572
3q	94,987,545	199,501,827	104,514,282	13q	17,918,001	114,142,980	96,224,979
4p	1	49,354,874	49,354,873	14q	18,070,001	106,368,585	88,298,584
4qA	52,354,875	75,641,303	23,286,428	15q	18,260,001	100,338,915	82,078,914
4qB	75,671,304	167,795,054	92,123,750	16p	1	35,143,302	35,143,301
4qC	167,825,055	191,273,063	23,448,008	16q	44,943,303	88,827,254	43,883,951
5p	1	46,441,398	46,441,397	17p	1	22,187,133	22,187,132
5qA	49,441,399	97,589,886	48,148,487	17q	22,287,134	78,774,742	56,487,608
5qB	97,612,887	180,857,866	83,244,979	18p	1	15,400,898	15,400,897
6p	1	58,888,125	58,888,124	18q	16,764,897	76,117,153	59,352,256
6qA	61,938,126	95,737,264	33,799,138	19pq	1	63,811,651	63,811,650
6qB	95,937,265	170,899,992	74,962,727	20pq	1	62,435,964	62,435,963
7p	1	58,058,273	58,058,272	21pq	1	46,944,323	46,944,322
7q	61,058,274	158,821,424	97,763,150	22q	14,430,001	49,691,432	35,261,431
8p	1	43,958,052	43,958,051				
8q	46,958,053	146,274,826	99,316,773				