Imputationを利用したジェノタイピングデータの推測

Imputationアルゴリズムは、連鎖不平衡（LD）ハプロタイプブロックにおけるSNPの相関を利用することで、コンテンツが異なるマーカーセット間でのジェノタイピングデータの推測を可能にします。

はじめに

ヒトの遺伝的多型のカタログは、国際HapMapプロジェクトなどの共同研究に伴い、この数年間で急速に拡大しています。1000人ゲノムプロジェクトなどの取り組みで、シーケンスデータがかつてないほどのペースで公共データベースに蓄積され続けている現在、ヒトの遺伝的多様性は今後もペースをおとすことなく発見が続くと考えられています。

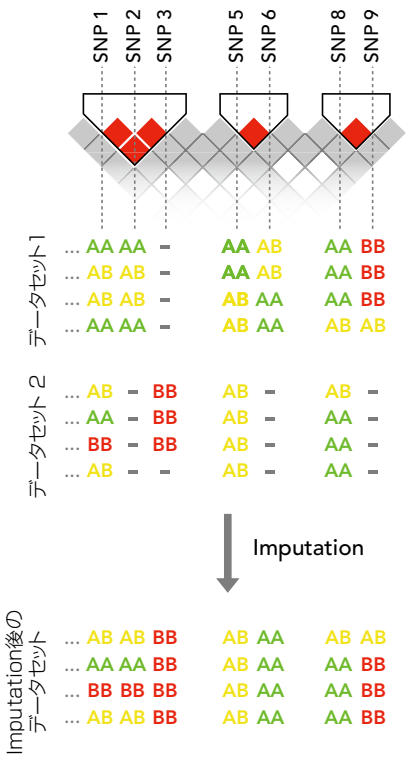
このような知識の拡大に伴い、この数年間で高スループットなジェノタイピング法の改良が急速に進みました。現在では、1回のマイクロアレイ実験で約500万データポイントのアッセイが可能です。ヒトの多型と疾患との関連についての研究が長期間にわたる一方で、マイクロアレイのテクノロジーや製品は急速に進化しています。そのため遺伝研究の過程では、最新のコンテンツを利用して研究をおこなうために、様々な世代のマイクロアレイを用いたデータが収集されています。

マーカーセットが異なる複数の種類のアレイを用いてデータを収集した場合、いくつかのマーカーは一部のアレイのみでアッセイされることになります。そのため、これらのマーカーについては、そのマーカーが搭載されているアレイで直接ジェノタイプしたサンプル数しか関連解析に使用できません。あるマーカーの解析に用いるサンプル数が限定されると、その解析における真の関連性の検出力も制限されてしまいます。しかし、近年のインフォマティクスの進歩により、2つのジェノタイピングアレイ間で共通していないマーカーのジェノタイプを補完するImputationのためのアルゴリズムが利用できるようになりました。Imputationを行うことで各マーカーにおけるサンプルサイズは増加し、その数はアレイの種類にかかわらずジェノタイピングされた個体の総数となります（図1）。

利用可能なソフトウェア

Imputationにはいくつかのフリーソフトウェアプログラムがあり、広く用いられています。いずれもUnix/Linuxベースのシステムで動作するコマンドラインプログラムです。最も一般的に用いられている4つのプログラムを表1に示します。ユーザーガイドと使用手順は、各プログラムのウェブサイトから入手できます。ダウンロードと使用に関する説明は、各プログラムの付属文書をご参照ください。

図1:IMPUTATIONの概要



SNP1～9は連鎖不平衡値が高い3つのブロックを構成しています。SNP間の赤い◆は連鎖不平衡を示します。データセット1と2は、SNP1～9において2つの異なるアレイを用いてジェノタイピングされた合計8データです。Imputation後のデータセットでは、データセット2で欠けていたデータをImputationして得られた推定ジェノタイプを含む、すべてのSNPローカスのジェノタイプが示されています。例えば、SNP2はデータセット1ではジェノタイピングされていますが、データセット2ではジェノタイピングされていません。SNP1～3は強い連鎖不平衡を示すため、データセット2ではデータセット1に示すジェノタイプに基づいて、SNP2における個々のジェノタイプを推定することができます。

IMPUTATIONに必要なシステム

Imputationでは高度な計算処理を必要とします。Imputationを行う場合、通常はUnixまたはLinuxベースの大規模クラスターにアクセスします。これらは必要に応じて所属研究機関のITまたはバイオインフォマティクス部門を通じて利用します。より速く解析を行うには、同時に複数の計算ノードにアクセスできるクラスターが必要です。

Imputationを染色体ごとに行う、または一度に数百サンプルのみを処理することにより、システム要件を緩和できます。このように分割したデータセットは、データ解析前に統合することが可能です。時間や計算能力が限られている場合は、1つの染色体またはターゲット領域のみに絞り込むことで、Imputationに必要なリソースを減らすことができます。

IMPUTATIONの計画にあたり考慮すべき事項

リファレンスとなる人種集団

高密度にジェノタイピングされたリファレンスとなるデータセットは、異なるアレイから得られたデータを整列させてImputationを行うための足場として利用できます。リファレンスデータは高密度のマーカーセットを提供するもので、それに付随するマイナーアレル頻度(MAF)と連鎖不平衡の情報を、データセット間の補完に用いることが可能です。リファレンスとなる人種集団は実験で用いたサンプルを代表するものでなければなりません[9]。例えば、白人から実験データを収集した場合は、白人の基準サンプル(例:HapMap　CEUサンプル)を用いる必要があります。同様に、混血や別の人種のサンプルであれば、それに見合うサンプルをリファレンスとして用いる必要があります。Huangらは、世界の多様な集団に対して補完を行う際のCEU/CHB/JPT/YRI　HapMapサンプルの最適比を提示しています[10]。

ストランドの一貫性

データセットを統合する際、2つのデータセットから得られたジェノタイプは常に同一ストランド(例:順鎖または「+」鎖)に由来していなければなりません。この一貫性が崩れるとデータの統合が不可能になったり、A/TおよびC/G SNPに不可解な反転が生じて誤った結果が得られたりします[7]。

いくつかの情報ソースから提供されているHapMapリファレンスデータはforwardストランドとなっています(後述のリファレンスデータセットのセクションに、2つ情報ソースのウェブサイトを例示しました)。イルミナでは、Infinium® HD HumanOmni1-Quad BeadChip以降のすべての製品について、ストランドアノテーションファイルを提供しています(テクニカルサポートにお問い合わせください)。これらのストランドアノテーションファイルを用いると、reverseストランド上でアッセイされたマーカーを同定できます。リファレンスデータとの統合およびImputationを行う前に、PLINKなどのプログラムを利用して、reverseストランドマーカーを反転させることが可能です。

ただし、HapMapデータにはストランドエラーが疑われる部分があるので、一部のマーカー(通常は全ゲノムデータセットのうち数千個以下)についてはストランドの方向が矛盾することになると予想しておく必要があります。このようなSNPは実験データから除去した上でImputationを行うことを推奨します。

データのクオリティチェック

ジェノタイプのImputationを行う前に、実験データセットから得られたジェノタイプについて、基本的なデータクオリティチェックを行っておくことをお奨めします。一般的には以下の項目を除去します[7]。

- **コールレートが低いマーカー**
- **Hardy-Weinberg平衡からの大きな逸脱が認められるもの**
- **繰り返し実験サンプル間での多数の不一致**
- **メンデル則からの逸脱が認められるもの**
- **MAFが非常に低いマーカー**

これらの指標に対する最適なカットオフ値は一定ではないので、それぞれの研究に応じて適切なスコアを用いる必要があります。

「確実な」ジェノタイプコールの信頼度閾値

異なるソースに由来するデータセットを統合した解析を行うためにImputationを用いる場合は、2つのデータセット間で共通していないマーカーについてのジェノタイプコールを行うことが目的となります。それぞれの非共通マーカーに対してジェノタイプをコールし、以後の関連解析に用いることができます。

表1:一般的に用いられているIMPUTATIONソフトウェアパッケージ

ソフトウェア名	研究機関	URL
MACH	ミシガン大学[1,2]	http://www.sph.umich.edu/csg/abecasis/MaCH/tour/imputation.html
BEAGLE	オークランド大学[3]	http://www.stat.auckland.ac.nz/~bbrowning/beagle/beagle.html
IMPUTE	オックスフォード大学[4,5]	http://mathgen.stats.ox.ac.uk/impute/impute.html
PLINK	マサチューセッツ総合病院／ブロード研究所[6]	http://pngu.mgh.harvard.edu/~purcell/plink/

Imputationでは、考えられる3つのジェノタイプのそれぞれについて確率を計算し、各位置で最も可能性が高いジェノタイプに基づいてコールを行います[9]。確実なジェノタイプコールが行われた場合、そのコールには正しいジェノタイプがコールされた確率に対応する信頼度スコアが付随します。例えば、ジェノタイプがAAである確率が95%、ABである確率が3%であれば、真のジェノタイプがAAであるという高い確率がAAの信頼度スコアに反映されることになります。AAである確率が40%、ABである確率が30%であれば、確実なジェノタイプコールが得られず、低い信頼度スコアに反映されます。信頼度スコアに基づく厳格なカットオフ値を設定することで、以後の関連解析におけるImputationエラーの確率が低下します[9]。信頼度スコアの解釈に関する詳細は、それぞれのImputationソフトウェアの付属文書を参照してください。

IMPUTATIONの精度

Imputationのエラーを最少にするには、厳格な信頼度スコアカットオフ値の設定以外にも、いくつかの対策を取ることができます。これらの注意点を考慮することにより、Imputationエラーに起因する不正確なデータ解釈の可能性が低下します。

Imputationは連鎖不平衡に基づくため、完全に独立したゲノム領域については予測されません。隣接マーカーの関連解析の結果は、補完マーカーと比較して同程度の関連性を示している必要があります。したがって、補完マーカーの関連性が、直接ジェノタイピングされた周辺のマーカーと統計的に大幅に異なっている場合には、注意して取り扱い、慎重に検討しなければなりません。リファレンスデータセットと統合される複数のデータセットにおいて大量のジェノタイピングデータが欠損しているという場合には例外もありえます。例えば、あるSNPに関するデータが50%のサンプルで得られず、隣接するSNPについても50%のサンプルでデータが得られていない場合、これらのマーカーについてはImputationにより全データセット間でほぼ100%のジェノタイプが得られる可能性があります。このようにImputationを行うと、各SNPにおけるジェノタイピング結果をもつサンプル数が増加し、検出力が劇的に高まり、個々のデータセットだけでは見出せなかった新たな関連性を発見できる可能性があるのです。

Imputationソフトウェアはいずれも「ゲノムの連鎖不平衡」という同一の基本的現象を利用したものですが、各ソフトウェアパッケージで採用されているアルゴリズムは異なります。同様に、各パッケージの長所や短所も異なります。したがって、複数のソフトウェアパッケージを用いて結果を比較し、大きな不一致の有無を検討されることをお勧めします。

Imputationによるジェノタイプコールの後は常に若干のエラーが残存するので、上位の関連性シグナルにImputationデータが含まれている場合には、比較的少数のジェノタイピングを行ってジェノタイプコールを確認することを推奨します。方法としては、数百例を対象として少数のSNPのジェノタイピングを行い、これらの個々のサンプルにおいて対象となるマーカーのImputation精度を確認することがあげられます。この精度は全データセットに対して適用することができます。

IMPUTATIONを用いたジェノタイプエラーチェック

Imputationを用いてジェノタイピングエラーを発見することも可能です。PLINKソフトウェアパッケージには、ゲノム上のジェノタイピングされたマーカーを1つずつ外してデータを再度Imputationし、リアルタイムで関連解析を行うという「drop-one」オプションがあります[6]。直接ジェノタイピングデータを用いて計算した関連性に関する統計量は、Imputationデータを用いた場合とよく一致するはずです。Imputationデータの信頼度が高い場合、大きな不一致は系統的バイアスの原因となりうるジェノタイピングエラーの存在を示すと考えられます。

マイナーアレル頻度とIMPUTATION効率

Imputationはゲノムに存在する連鎖不平衡とハプロタイプブロック構造に依存するので、多型がまれであるほど、信頼度の高いジェノタイプImputationが困難になります。コモンSNP（MAF 5%以上）はImputationが可能であり、1000人ゲノムプロジェクトの初期の報告では、MAFが1.5%のレベルであっても大半の多型マーカーについてはImputationが可能だとされています（未発表データ）。イルミナの従来のWhole-Genome Genotyping BeadChipはMAF 5%以上のtag SNPを用いて構成されていますので、これらのアレイから得られたデータはチップ間でのImputationに極めて適しています[5]。

リファレンスデータセット

多くの研究において、独立した60例からなる標準HapMapパネルがリファレンスデータセットとして採用され、成果が得られています[3]。Imputationソフトウェアでの使用に適した形式のHapMapリファレンスデータセットは、以下をはじめとする多くのウェブサイトからダウンロードできます。

- http://mathgen.stats.ox.ac.uk/impute/impute.html
- http://pngu.mgh.harvard.edu/~purcell/plink/res.shtml

IMPUTATIONの例を示す文献

2008-2009年に発表された多くの大規模メタ解析は、近年のImputation法の進歩により可能となったものです。これらの研究では、多くの異なるプラットフォームや研究機関で収集されたデータを統合し、Imputationを用いてSNPコンテンツの違いに起因する欠損データを補いました。これらのメタ解析には数万ものサンプルが含まれますが、Imputationを用いることで検出力が劇的に高まり、多数の関連ローカスがさらに見出されました。Imputationの利用例については、以下の参考文献のセクションをご参照ください。

まとめ

Imputationは、異なるマーカーセットを用いて作成されたジェノタイピングデータセットを統合する際に、最大限の情報を利用可能にするための重要かつ有益な方法です。リファレンスデータセットにおける連鎖不平衡とジェノタイプ間の相関性を利用して、データセット中の欠損マーカーのジェノタイプを確実に推定できます。Imputationを行うためのいくつかのソフトウェアパッケージが利用可能です。多くの参考文献に、Imputationを用いた統合解析の例が記載されています。

参考文献（IMPUTATIONアルゴリズム）

(1) Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40:161-9.

(2) Sanna S, Jackson AU, Nagaraja R, Willer CJ, Chen WM, et al. (2008) Common variants in the GDF5-UQCC region are associated with variation in human height. Nat Genet 40:198–203.

(3) Browning BL and Browning SR (2009) A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84:210–223.

(4) Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies via imputation of genotypes. Nat Genet 39: 906-913.

(5) Howie B, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics 5(6): e100052.

(6) Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet, 81.

(7) de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, et al. (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet. 17(R2): R122–8.

(8) Pei YF, Li J, Zhang L, Papasian CJ, Deng HW. (2008) Analyses and comparison of accuracy of different genotype imputation methods. PLoS ONE 3(10): e3551.

(9) Guan Y, Stephens M. (2008) Practical issues in imputation-based association mapping. PLoS Genet 4(12): e1000279.

(10)Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, et al. (2009) Genotype-imputation accuracy across worldwide human populations. Am J Hum Genet. 84(2): 235–50.

(11)Spencer CC, Su Z, Donnelly P, Marchini J. (2009) Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. 5(5): e1000477.

(12)Padilla MA, Divers J, Vaughan LK, Allison DB, Tiwari HK. (2009) Multiple imputation to correct for measurement error in admixture estimates in genetic structured association testing. Hum Hered. 68(1): 65–72.

(13)Becker T, Flaquer A, Brockschmidt FF, Herold C, Steffens M. (2009) Evaluation of potential power gain with imputed genotypes in genome-wide association studies.Hum Hered. 68(1): 23–34.

参考文献（IMPUTATIONの利用例）

(14)Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40:161–169.

(15)Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, et al. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 40:638-645.

(16)Barrett JC, Hansoul S, Nicolae DL, Cho J, Duerr RH, et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40:955–962.

(17)Willer CJ, Speliotes EK, Loos RJF, Li S, Lindgren CM, et al. (2008) Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat Genet 41:25–34.

(18)Newton-Cheh C, Eijgelsheim M, Rice KM, deBakker PIW, Yin X, et al. (2009) Common variants at ten loci influence QT interval duration in the QTGEN Study. Nat Genet 41:399–406.

(19)Cho YS, Go MJ, Kim YJ, Heo JY, Oh JH, et al. (2009) A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nat Genet 41:527–534.

(20)Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, et al. (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet May 10 [ePub ahead of print].

(21)Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, et al. (2009) Genome-wide association study identifies eight loci associated with blood pressure. Nat Genet 41:666–676.

(22)De Jager PL, Jia X, Wang J, deBakker PI, Ottoboni L, Aggarwal NT, et al. (2009) Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. Nat Genet 41:776–782.

より詳しい情報

イルミナのDNA解析ツールについてさらに詳しくお知りになりたい場合は、弊社ウェブサイトをご覧いただくか、弊社までお問い合わせください。

イルミナ株式会社

〒108-0014

東京都港区芝5-36-7　三田ベルジュビル22階

Tel (03)4578-2800　Fax (03)4578-2810

www.illuminakk.co.jp

代理店

本製品の使用目的は研究に限定されます。

Pub. No. 370-2009-014 01 oct 09