जीनोम वाइड एसोसिएशन स्टडीज (जी डब्ल्यू ए एस) में इंप्यूटेशन तकनीकों के मुद्दे एवं चुनौतियां: एक समीक्षा

राहुल बनर्जी[1], भारती[1], शबाना बेगम[2], पंकज दास[1], तौकीर अहमद[1]

सारांश

जीनोम–वाइड एसोसिएशन स्टडी (GWAS) में बीमारी से जुड़ी जेनेटिक की खोज हेतु व्यक्तियों के डीएनए मार्कर्स को स्कैन किया जाता है। जब नए जेनेटिक संघों की पहचान होती है, तो इस जानकारी का उपयोग बीमारी की पहचान, इलाज और रोकथाम हेतु बेहतर रणनीतियां बनाने में किया जाता है। इम्प्यूटेशन जेनेटिक अध्ययनों में अनटाइप्ड जेनोटाइप का पूर्वानुमान किया जाता है जब डेटा गुणवत्ता, लागत, डिजाइन समस्याओं के कारण अनुपलब्ध होता है। यह एक प्रमाणित सांख्यिकी तकनीक है जिसका उपयोग घनी जेनोटाइप जाँच पैनल से हैप्लोटाइप सेगमेंट उधारण करके अनदेखे जेनोटाइप का आकलन करने हेतु किया जाता है। जेनोटाइप इम्प्यूटेशन जीनोम–व्यापक संघटन स्कैन के विश्लेषण करने में महत्वपूर्ण है। इस शोध पत्र में संक्षेप में मिसिंग डेटा समस्याओं और विभिन्न इम्प्यूटेशन विधियों को दर्शाया गया है।

शब्द कुंजी: बीगल, फास्फेस, जीनोम वाइड एसोसिएशन स्टडीज, इंप्यूटेशन के तरीके, इम्प्यूट, मैच, मिसिंग मैकेनिज्म।

Issues and Challenges of Imputation Techniques in Genome Wide Association Studies (GWAS): A Review

Rahul Banerjee[1], Bharti[1], Shbana Begum[2], Pankaj Das[1], Tauqueer Ahmad[1]

10.18805/BKAP597

ABSTRACT

A genome-wide association study (GWAS) rapidly scans DNA markers in many individuals to find genetic links to diseases. New findings aid in disease detection, treatment and prevention. Imputation predicts untyped genotypes in genetic studies when data is missing due to quality, cost, or design issues. It's a proven statistical technique for estimating unobserved genotypes by borrowing haplotype segments from a densely genotyped reference panel. This allows estimation and testing of associations at unassayed variants. Genotype imputation is vital in analyzing genome-wide association scans, helping geneticists evaluate evidence for association at untyped genetic markers. This summary outlines missing data issues and various imputation methods.

Key words: BEAGLE, fastPHASE, Genome wide association studies, Imputation methods, IMPUTE, MACH, Missing mechanisms.

आंकड़ों में, मिसिंग डेटा तब होता है जब किसी अवलोकन में चर के लिए कोई डेटा मान संग्रहीत नहीं किया जाता है। मिसिंग डेटा एक सामान्य घटना है और अनुसंधान की शुरुआत के बाद से शोधकर्ताओं को चुनौती दी है, विशेष रूप से अनुदैर्ध्य अनुसंधान के लिए, जिसमें एक ही व्यक्ति पर बहुत सारे आँकड़े होते है। शोधकर्ताओं द्वारा उपयोग की जाने वाली प्रक्रियाओं को मुख्य रूप से बीसवीं शताब्दी में विकसित किया गया था जिसे संपूर्ण डेटा के लिए विकसित किया गया है। मिसिंग डेटा सांख्यिकीय अनुमानों की वैधता पर प्रतिकूल प्रभाव डाल सकते हैं।

[1]ICAR-Indian Agricultural Statistics Research Institute, Library Avenue, Pusa-110 012, New Delhi, India.
[2]ICAR-National Institute for Plant Biotechnology, LBS Centre, Pusa-110 012, New Delhi, India.

Corresponding Author: Rahul Banerjee, ICAR-Indian Agricultural Statistics Research Institute, Library Avenue, Pusa-110 012, New Delhi, India. Email: rahuliasri@gmail.com

How to cite this article: Banerjee, R., Bharti, Begum, S., Das, P. and Ahmad, T. (2023). Issues and Challenges of Imputation Techniques in Genome Wide Association Studies (GWAS): A Review. Bhartiya Krishi Anusandhan Patrika. 38(3): 193-202. doi: 10.18805/BKAP597.

Submitted: 21-09-2022 **Accepted:** 23-09-2023 **Online:** 13-10-2023

जीनोम वाइड एसोसिएशन स्टडीज (GWAS)

आनुवंशिकी में, जीनोम–वाइड एसोसिएशन स्टडी (GWA स्टडी, GWAS), जिसे संपूर्ण जीनोम एसोसिएशन स्टडी (WGA स्टडी, WGAS) के रूप में भी जाना जाता है, जोकि विभिन्न व्यक्तियों में आनुवंशिक वेरिएंट के जीनोम–वाइड सेट का एक परीक्षण है, यह देखने के लिए किया जाता है, कि क्या कोई प्रकार है एक विशेषता के साथ जुड़ा हुआ है? GWAS आमतौर पर एकल–न्यूक्लियोटाइड बहुरूपताओं (एसएनपी) और प्रमुख मानव रोगों के लक्षणों के बीच संघो पर ध्यान केंद्रित करते है, लेकिन समान रूप से किसी भी अन्य जीव पर लागू किया जा सकता है। जीनोम–वाइड एसोसिएशन स्टडीज वैज्ञानिकों के लिए मानव रोग में शामिल जीन की पहचान करने का एक अपेक्षाकृत नया तरीका है। यह विधि जीनोम को छोटे बदलावों को खोजती है, जिन्हें सिंगल न्यूक्लियोटाइड पॉलीमॉर्फिज्म या एसएनपी कहा जाता है, जो बिना बीमारी वाले लोगों की तुलना में किसी विशेष बीमारी वाले लोगों में अधिक बार होते हैं। प्रत्येक अध्ययन एक ही समय में सैकड़ों या हजारों एसएनपी देख सकता है। शोधकर्ता इस प्रकार के अध्ययन के डेटा का उपयोग उन जीनों को इंगित करने के लिए करते हैं जो किसी व्यक्ति के एक निश्चित बीमारी के विकास में योगदान कर सकते हैं। जब मानव डेटा पर लागू किया जाता है, तो GWA स्टडीज प्रतिभागियों के डीएनए की तुलना किसी विशेष लक्षण या बीमारी के लिए अलग–अलग फेनोटाइप वाले करते हैं। ये प्रतिभागी एक बीमारी वाले लोग एवं बिना (नियंत्रण) के समान लोग हो सकते हैं, या वे किसी विशेष लक्षणों के लिए अलग–अलग फेनोटाइप वाले लोग हो सकते हैं, उदाहरण के लिए रक्तचाप। इस दृष्टिकोण को फेनोटाइप–प्रथम के रूप में जाना जाता है, जिसमें प्रतिभागियों को पहले उनके नैदानिक अभिव्यक्तियों द्वारा वर्गीकृत किया जाता है, जैसा कि जीनोटाइप–प्रथम के विपरीत होता है। प्रत्येक व्यक्ति डीएनए का एक नमूना देता है, जिसमें से लाखों आनुवंशिक रूपों को एसएनपी सारणियों का उपयोग करके पढ़ा जाता है। यदि बीमारी वाले लोगों में एक एलील अधिक पाया जाता है, तो इस एलील को बीमारी से जुड़ा हुआ कहा जाता है। कोई भी दो मानव जीनोम लाखों अलग–अलग तरीकों से भिन्न होते हैं। जीनोम (एसएनपी) के अलग–अलग न्यूक्लियोटाइड में छोटे बदलाव होते हैं और साथ ही कई बड़े बदलाव होते हैं, जैसे विलोपन, सम्मिलन और प्रतिलिपि संख्या (कॉपी नंबर) भिन्नताएं। वर्ष 2000 के आसपास, जीडब्ल्यूए अध्ययनों की शुरुआत से पहले, जांच का प्राथमिक तरीका परिवारों में अनुवांशिक संबंधों के वंशानुक्रम अध्ययन के माध्यम से किया जाता था। यह दृष्टिकोण एकल जीन विकारों के लिए अत्यधिक उपयोगी साबित हुआ था। हालांकि, जटिल बीमारियों के लिए आनुवंशिक लिंकेज अध्ययन के परिणामों को पुनः पेश करना मुश्किल साबित होता है। लिंकेज अध्ययन का एक सुझाया गया विकल्प आनुवंशिक एसोसिएशन स्टडी है। यह अध्ययन जांचता है कि क्या वे व्यक्तियाँ जिनके पास एलील रुचि से संबंधित भौतिक गुण होते हैं, उन्हें आनुवंशिक रूपांतर के संदर्भ में अपेक्षित से अधिक बार देखा जाता है।

पद्धति

जी डब्ल्यू ए अध्ययनों का सबसे आम दृष्टिकोण केस–कंट्रोल सेटअप है, जो व्यक्तियों के दो बड़े समूहों, एक स्वस्थ नियंत्रण समूह और एक बीमारी से प्रभावित, केस समूह की तुलना करता है। एसएनपी की सटीक संख्या जीनोटाइपिंग तकनीक पर निर्भर करती है, लेकिन आमतौर पर यह एक मिलियन या अधिक भी हो सकती है। इनमें से प्रत्येक एसएनपी की जांच की जाती है, और देखा जाता है कि क्या एलील आवृत्ति उपचार एवं नियंत्रण समूह के बीच महत्वपूर्ण रूप से बदल जाती है। ऑड्स अनुपात, दो ऑड्स का अनुपात है, जो GWA अध्ययनों के संदर्भ में एक विशिष्ट एलील वाले व्यक्तियों के लिए रोग की संभावना है और उन व्यक्तियों के लिए रोग की संभावना है जिनके पास समान एलील नहीं है।

गणितीय रूप से,

माना A कोई घटना है तब, किसी घटना के घटित होने की संभावना = P(घटना घटती है)/P(घटना घटित नहीं होती)=

$$\frac{P(A)}{P(A^c)} = \frac{P(A)}{1-P(A^c)}$$

एक उदाहरण लेते हैं:

	एलील काउंट	
	G	T
केसेस	a	b
नियंत्रण	c	d

नमूने में सभी G एलील पर विचार करें, और यादृच्छिक रूप से एक चुनें। एक मामले में G एलील होने की संभावना: $:\frac{a}{c}$। नमूने में सभी T एलील पर विचार करें, और यादृच्छिक रूप से एक चुनें। एक मामले में एक टी T एलील होने की संभावना: $\frac{b}{d}$।

ऑड्स अनुपात (OR): = ad/bc

व्याख्या इस प्रकार कही जा सकती है:

OR= प्रत्येक अतिरिक्त G एलील के लिए केस होने की संभावना में वृद्धि।

OR=1: जीनोटाइप और रोग के बीच कोई संबंध नहीं।

OR>1: G एलील से बीमारी का खतरा बढ़ जाता है।

OR<1: T एलील से बीमारी का खतरा बढ़ जाता है।

जब केस ग्रुप में एलील आवृत्ति नियंत्रण ग्रुप की तुलना में बहुत अधिक होती है, तो ऑड्स अनुपात 1 से अधिक होता है, और कम एलील आवृत्ति के लिए इसके विपरीत। इसके अतिरिक्त, ऑड्स अनुपात के महत्व के लिए P–मान की गणना आम तौर पर एक साधारण काई–स्क्वेर्ड (χ^2) टेस्ट का उपयोग करके की जाती है। जीडब्ल्यूए अध्ययन का उद्देश्य 1 से काफी भिन्न ऑड्स अनुपातों का पता लगाना है क्योंकि इससे पता चलता है कि एसएनपी बीमारी से जुड़ा है या नहीं। जीनोटाइप इंप्यूटेशन सांख्यिकीय विधियों द्वारा किया जाता है जो GWAS डेटा को हैप्लोटाइप के संदर्भ पैनल के साथ जोड़ते हैं। जीनोटाइप इंप्यूटेशन के लिए मौजूदा सॉफ्टवेयर पैकेज में IMPUTE2 और MacH शामिल हैं।

मिसिंग डेटा के स्रोत

सर्वेक्षण अनुसंधान में मिसिंग डेटा के तीन मुख्य स्रोत हैं: नॉन–कवरेज, कुल गैर–प्रतिक्रिया (नॉन रेस्पोंस), और आइटम गैर–प्रतिक्रिया (ग्रोव्स और अन्य 2004)। नॉन–कवरेज तब होता है जब कुछ समष्टि के नमूने में चुने जाने की कोई संभावना नहीं होती है। मिसिंग डेटा गैर–प्रतिक्रिया से आता है जब कोई प्रतिवादी सर्वेक्षण पर किसी भी आइटम का जवाब देने से इनकार करता है। आइटम गैर–प्रतिक्रिया तब होती है जब सर्वेक्षण पर केवल आंशिक आइटम पूरा किया जाता है। लापता डेटा के इन तीन स्रोतों में, गैर–कवरेज और कुल गैर–प्रतिक्रिया को उपयुक्त नमूना भार का उपयोग करके संबोधित किया जा सकता है जो नमूना को लक्षित समष्टि का सटीक रूप से प्रतिनिधित्व करने के लिए डिजाइन किया गया है (चीमा, 2014)। हालांकि, आइटम गैर–प्रतिक्रिया के कारण मिसिंग मान वजन का उपयोग करके तय नहीं किए जा सकते हैं। आइटम गैर–प्रतिक्रिया वाले मामलों को सूचीवार हटाने से कुछ मूल्यवान जानकारी का नुकसान होता है और संभावित रूप से बायस पैदा होती है। मिसिंग डेटा के साथ एक मुख्य चिंता यह है कि क्या पूर्ण डेटा वाला नमूना अभी भी लक्षित आबादी समष्टि का प्रतिनिधि है (रोथ, 1994)।

मिसिंग तंत्र

मिसिंग डेटा तंत्र को आम तौर पर तीन समूहों में वर्गीकृत किया जाता है (रुबिन, 1976), यादृच्छिक रूप से पूरी तरह से गायब (MCAR), यादृच्छिक रूप से गायब (MAR) एवं यादृच्छिक रूप से गायब नहीं (MNAR)। यदि जिन विषयों में आँकडे गुम है, वे विषयों के पूरे नमूने का एक यादृच्छिक उपसमुच्चय हैं, तो मिसिंग डेटा को पूरी तरह से यादृच्छिक कहा जाता है (MCAR)। केवल पूर्ण डेटा का उपयोग करने से निष्पक्ष परिणाम मिल सकता है। MAR तब होता है जब गुम होना प्रेक्षित डेटा पर निर्भर नहीं करता है। डेटा MNAR तब होते हैं जब किसी चर पर डेटा गुम होने की संभावना उस चर के मान पर निर्भर करती है।

आनुवंशिक महामारी विज्ञान में मिसिंग डेटा

आनुवंशिक महामारी विज्ञान के अध्ययन में, डेटा की गुणवत्ता, लागत दक्षता या तकनीकी डिजाइन के कारणों के विश्लेषण के लिए विशेष मार्करों के जीनोटाइप अनुपलब्ध होने पर मिसिंग डेटा समस्याएं उत्पन्न होती हैं। सबसे मुख्य जीनोटाइप इंप्यूटेशन विधियों में IMPUTE, FastPHASE, MacH और BEAGLE शामिल हैं (मार्चिनी और होवी, 2010)। मानव आनुवंशिक महामारी विज्ञान का उद्देश्य विशिष्ट फेनोटाइप से संबंधित रोग या रोग–संबंधी लक्षण पर आनुवंशिक रूपांतरों की पहचान करना है। जीनोटाइपिंग अध्ययन हमेशा चयनित मार्कर सेटों तक ही सीमित रहा है, मुख्य रूप से एकल न्यूक्लियोटाइड बहुरूपता (एसएनपी) जो व्यावसायिक रूप से उपलब्ध माइक्रोएरे पर स्वरूपित होते हैं। लिंकेज असाम्यवस्था पर आधारित जीनोम–वाइड एसोसिएशन स्टडीज (जीडब्ल्यूएएस), केवल एसएनपी की एक छोटी संख्या का उपयोग करता है जो किसी दी गई समष्टि में 80% आनुवंशिक भिन्नता का विवरण करती है (क्रॉस्कैक, 2015)। 5 जीनोटाइप इंप्यूटेशन विधियाँ वैज्ञानिकों को मिसिंग डेटा समस्या को संबोधित करने में मदद कर सकती हैं (मार्चिनी एंड होवी, 2010)।

इम्प्यूटेशन विधियाँ

महामारी विज्ञान के अध्ययन के लिए विभिन्न जीनोटाइप इंप्यूटेशन विधियों का प्रस्ताव किया गया है। उन जीनोटाइप इंप्यूटेशन विधियों के बीच मुख्य अंतर यह है कि एक एसएनपी के सशर्त जीनोटाइप वितरण को कैसे परिभाषित और उपयोग किया जाता है। मान लें कि हमारे पास L डायलेलिक ऑटोसोमल एसएनपी में डेटा है और प्रत्येक एसएनपी में दो

एलील को 0 और 1 से कोडित किया गया है। H, इन L एसएनपी में N हैप्लोटाइप के एक सेट को दर्शाता है जबकि G, L एसएनपी में जीनोटाइप डेटा का सेट है। जीनोटाइप इंप्यूटेशन का उद्देश्य उन एसएनपी के जीनोटाइप की भविष्यवाणी करना है जिन्हें अध्ययन नमूने में जीनोटाइप नहीं किया गया है। यहाँ उल्लिखित सभी इंप्यूटेशन विधियां मानती हैं कि मिसिंग डेटा MAR है एवं सारे HMM आधारित इंप्यूटेशन विधियाँ हैं, यह संदर्भ पैनल पर सशर्त, अध्ययन में लिए गए प्रत्येक व्यक्ति के मिसिंग जीनोटाइप का अनुमान लगाते हैं। ऐसा करने से संदर्भ पैनल का उपयोग प्रत्येक व्यक्ति के मल्टीलोकस जीनोटाइप में चरण अनिश्चितता पर विशलेषणात्मक रूप से एकीकृत करने के लिए किया जाता है।

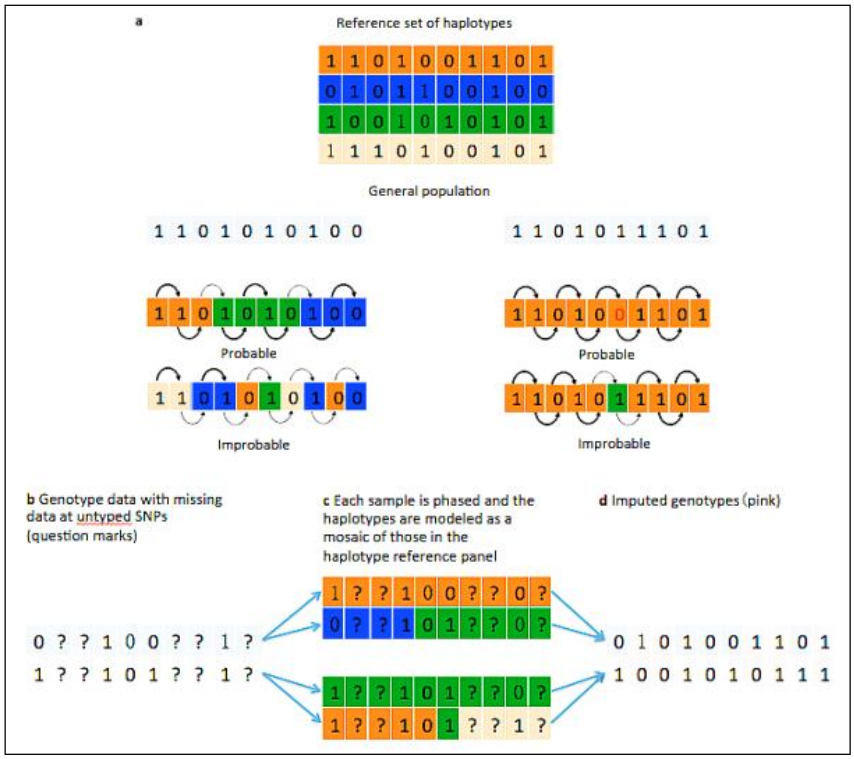
इम्प्यूट (IMPUTE)

मार्चिनी एट अल (2007) ने IMPUTE विधि का पहला संस्करण तैयार किया (चित्र 1)। IMPUTEv1 सहसंयोजी वृक्षों के अनुकरण के लिए हिडेन मार्कोव मॉडल (HMM) के विस्तार पर आधारित है। (फेयरनहेड एवं डोनेली, 2001) और लिंकेज डिसिपिलिब्रियम के मॉडलिंग और पुनर्संयोजन दरों के आकलन के लिए (ली एंड स्टीफेंस, 2003)। यह विधि प्रत्येक व्यक्ति के जीनोटाइप वेक्टर G_i के एचएमएम पर आधारित है जो N ज्ञात हैप्लोटाइप H के एक सेट पर सशर्त है। एक छिपे हुए मार्कोव मॉडल (एचएमएम) का रूप इस प्रकार होता है:

$$\Pr(G_i \mid H) = \sum_{Z_i^{(1)} Z_1^{(2)}} \Pr(G_i \mid Z_i^{(1)}, Z_1^{(2)}, H)\Pr(Z_i^{(1)}, Z_i^{(2)}, H)$$

जहां, $Z_i^{(1)} = \{Z_{i1}^{(1)},...,Z_{iL}^{(1)}\}$ और $Z_i^{(2)} = \{Z_{i1}^{(1)},...,Z_{iL}^{(2)}\}$L साइटों पर हिडेन स्टेट के दो क्रम हैं और $Z_i^{(j)} \in \{1,...N\}$ Z_i को संदर्भ पैनल H से हैप्लोटाइप की जोड़ी के रूप में माना जा सकता है जिसे जीनोटाइप वेक्टर G_i बनाने के लिए कॉपी किया जाता है। $\Pr(G_i^{(1)}, Z_i^{(2)}, H)$ यह पूर्व संभावना को परिभाषित करता है कि कैसे कॉपी किए गए हैप्लोटाइप्स की जोड़ी अनुक्रम के साथ बदलती है और एक मार्कोव श्रृंखला द्वारा परिभाषित की जाती है जिसमें स्विचिंग दर जीनोम भर में ठीक–पैमाने पर पुनर्संयोजन मानचित्र के अनुमान पर निर्भर करती है। मार्कोव श्रृंखला की प्रारंभिक अवस्था N^2 स्टेट पर एक समान है।

$$\Pr(G_i^{(1)}, Z_i^{(2)}, H) = \frac{1}{N^2}$$



चित्र 1: क्रावजाक (2015) पर आधारित इंप्यूट के साथ जीनोटाइप इंप्यूटेशन।

हम साइट / से /+1 तक श्रृंखला की संक्रमण संभावनाओं की गणना कर सकते हैं:

$$\Pr(\{Z_{il}^{(1)},Z_{il}^{(2)}\} \rightarrow \{Z_{i(l+1)}^{(1)},Z_{i(l+1)}^{(2)}\},H)$$
$$= \begin{cases} (e^{-\frac{\rho_l}{N}} + \frac{1-e\frac{\rho_l}{N}}{N})^2, Z_{il}^{(1)} = Z_{i(l+1)}^{(1)}, Z_{il}^{(2)} = Z_{i(l+1)}^{(2)} \\ (e^{-\frac{\rho_l}{N}} + \frac{1-e\frac{\rho_l}{N}}{N})(\frac{1-e\frac{\rho_l}{N}}{N}), Z_{il}^{(1)} = Z_{i(l+1)}^{(1)}, Z_{il}^{(2)} \neq Z_{i(l+1)}^{(2)}, Z_{il}^{(1)} \neq Z_{i(l+1)}^{(1)}, Z_{il}^{(2)} = Z_{i(l+1)}^{(2)} \\ (e^{-\frac{\rho_l}{N}} + \frac{1-e\frac{\rho_l}{N}}{N}), Z_{il}^{(1)} \neq Z_{i(l+1)}^{(1)}, Z_{il}^{(2)} \neq Z_{i(l+1)}^{(2)} \end{cases}$$

$\rho_t = 4N_e r_l$ और r_l साइटों / और /+1 के बीच प्रति पीढ़ी आनुवंशिक दूरी है। पूर्व वितरण के रूप में लिखा जा सकता है:

$$\Pr\left(Z_i^{(1)},Z_i^{(2)},H\right) = \Pr(Z_{il}^{(1)},Z_{il}^{(2)},H)\prod_{i=1}^{L-1}\Pr(\{Z_{il}^{(1)},Z_{il}^{(2)}\} \rightarrow \{Z_{i(l+1)}^{(1)},Z_{i(l+1)}^{(2)}\},H)$$

यह, $\Pr(G_t \mid Z_t^{(1)}, Z_t^{(2)}, H)$ परिभाषित करता है कि देखे गए जीनोटाइप कैसे करीब होंगे, लेकिन ठीक उसी तरह नहीं जैसे हैप्लोटाइप की नकल की जाती है।

इस चित्र में 0 और 1 एक संदर्भ एलील की उपस्थिति या अनुपस्थिति को दर्शाता है, इंप्यूटेशन बेस (हैप्लोटाइप का संदर्भ सेट) चरणबद्ध है और इसमें चार अलग–अलग हैप्लोटाइप शामिल हैं (ए) जनसंख्या हैप्लोटाइप को संदर्भ सेट से हैप्लोटाइप का मोजेक माना जाता है। संक्रमण संभावनाओं के साथ मार्कोव श्रृंखला मॉडल द्वारा परिभाषित, संबंधित हैप्लोटाइप वितरण जनसंख्या इतिहास और स्थानीय पुनर्संयोजन मानचित्र दोनों पर निर्भर करता है। बोल्ड तीर उच्च संक्रमण संभावनाओं को इंगित करते हैं; जबकि लगता है कि तीर कम संक्रमण संभावनाओं का प्रतिनिधित्व करते हैं। लाल का अर्थ है म्युटेशन। (बी) बिना टाइप किए गए एसएनपी (प्रश्न चिह्न) पर लापता डेटा के साथ जीनोटाइप डेटा को दर्शाता है। (सी) प्रत्येक नमूने को चरणबद्ध किया जाता है और हैप्लोटाइप को हैप्लोटाइप संदर्भ पैनल में मोजेक के रूप में तैयार किया जाता है। (डी) अध्ययन के नमूने में मिसिंग जीनोटाइप को संदर्भ सेट में मिलान करने वाले हैप्लोटाइप का उपयोग करके लगाया जाता है। सामान्य तौर पर, मिसिंग डेटा के साथ एक चरणबद्ध जीनोटाइप की संभावना का मूल्यांकन मोजेक हैप्लोटाइप के सभी संभावित जोड़े पर विचार करके किया जाता है जो देखे गए डेटा के साथ संगत होंगे। और सबसे संभावित जोड़ी गैर–टाइप किए गए एसएनपी पर सबसे संभावित जीनोटाइप निर्धारित करती है। IMPUTEv2 को ली एवं स्टीफंस (2003) द्वारा विकसित किया गया है, जो IMPUTEv1 की तुलना में ज्यादा अच्छा दृष्टिकोण है। प्रत्येक व्यक्ति के लिए एक अलग, विश्लेषणात्मक इम्प्यूट चरण करने के बजाय, IMPUTEv2 सभी व्यक्तियों को एक पुनरावृत्त ढांचे में एक साथ लगाता है। एसएनपी को अध्ययन के नमूने और संदर्भ पैनल (सेट टी) दोनों में जीनोटाइप किया जाता है सिर्फ संदर्भ पैनल (सेट यू) में जीनोटाइप किया जाता है, इस आधार पर एसएनपी को दो अलग–अलग सेटों में विभाजित किया जाता है। अध्ययन के नमूने (सेट टी) में एसएनपी पर हैप्लोटाइप का अनुमान पहले लगाया जाता है और फिर यू में एसएनपी में एलील को वर्तमान में अनुमानित हैप्लोटाइप पर सशर्त लगाया जाता है। विशेष रूप से, यह एक मार्कोव श्रृंखला मोंटे कार्लो (एमसीएमसी) एल्गोरिदम चलाता है जो दो बुनियादी चरणों के बीच वैकल्पिक होता है: 1) चरण सभी देखे गए जीनोटाइप और अध्ययन नमूने में एसएनपी में किसी भी छिटपुट रूप से लापता जीनोटाइप को लागू करें (सेट टी में एसएनपी)। किसी दिए गए एसएनपी में जीनोटाइप किए गए सभी व्यक्तियों के लिए पूल चरण की जानकारीय 2) पिछले चरण में अनुमानित प्रत्येक हैप्लोटाइप के लिए, गैर–टाइप किए गए एसएनपी (होवी और मार्चिनी, 2009) में लापता एलील्स को लागू करने के लिए संदर्भ पैनल का उपयोग करें। हैप्लोटाइप्स के लिए एक संभाव्यता वितरण को IMPUTEv1 के समान परिभाषित किया गया है, और दो हैप्लोटाइप्स में उच्चतम पश्च प्रायिकता है जो अध्ययन नमूने के लिए चुने गए हैं। IMPUTEv2 IMPUTEv1 की तुलना में बहुत तेज है क्योंकि इंप्यूटेशन चरण अगुणित प्रतिरूपण है,

IMPUTEv1 में O (N^2) से IMPUTEv2 में O (N) तक गणना समय को कम करता है (N संभावित हैप्लोटाइप की संख्या को दर्शाता है) (होवी, डोनेली और मार्चिनी, 2009; मार्चिनी और होवी, 2010)।

फास्टफैज (fast PHASE)

यह विधि एसएनपी हैप्लोटाइप्स पर आधारित है, जो समान हैप्लोटाइप्स (शीट एंड स्टीफेंस, 2006) वाले समूहों में क्लस्टर करते हैं, और इसे BIMBAM (सर्विन एंड स्टीफेंस, 2007) नामक एक एसोसिएशन–परीक्षण कार्यक्रम में लागू किया गया है। यह मॉडल सामान्य हैप्लोटाइप्स का प्रतिनिधित्व करने के लिए K अप्रकाशित अवस्थाओ के एक सेट को निर्दिष्ट करता है और प्रत्येक क्लस्टर (k^{th}) को साइट/पर प्रत्येक क्लस्टर में निहित हैप्लोटाइप्स के अंश के अनुपात में एक वजन (a_{ki}) सौंपा जाता है।

$$\sum_{k} a_{ki} = 1$$

प्रत्येक क्लस्टर में प्रत्येक साइट पर एलील 1 की आवृत्ति है। FastPHASE मॉडल जनसंख्या हैप्लोटाइप अवस्थाओ के बीच संक्रमण के साथ एक HMM के रूप में।

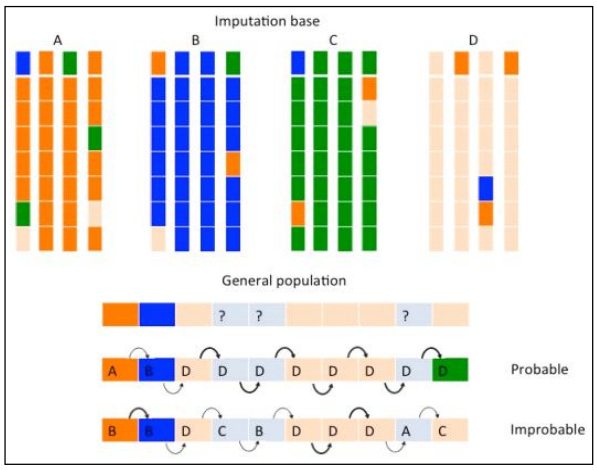
$$P(G_i | \alpha,\theta,r) = \sum_{z} P(G_i | Z_{i,} \theta)\ P | \alpha,r)$$

$Pr(G_i | Z,\theta)$ परिभाषित करता है कि देखे गए जीनोटाइप और अवस्था के बीच स्विच करने के मॉडल पैटर्न की कितनी संभावना है जो समूहों का प्रतिनिधित्व करते हैं न कि संदर्भ हैप्लोटाइप। और यहाँ, एक संभावना के रूप में लिखा जा सकता है।

$$L(G_i G | \alpha,\theta,r) = \prod P(G_i | \alpha,\theta,r) \prod P(H_i | \alpha,\theta,r)$$

मॉडल को फिट करने के लिए एक ईएम एल्गोरिदम का उपयोग किया जाता है, और पैरामीटर अनुमान पर सशर्त मिसिंग जीनोटाइप को लागू करने के लिए एक फॉरवर्ड–बैकवर्ड एल्गोरिदम का उपयोग किया जाता है। शोधकर्ताओं ने पाया कि अनुमानों के एक सेट के औसत से केवल एक अनुमान की तुलना में बहुत बेहतर परिणाम मिले। FastPHASE स्टेट के छोटे सेटों का उपयोग करता है जो कि यथोचित रूप से कम्प्यूटेशनल रूप से तेज होना चाहिए, हालांकि, यह लाभ आंशिक रूप से अमूर्त समूहों के साथ काम करके अधिक होता है जिसके लिए कई मापदंडों की आवश्यकता होती है।

जैसा कि चित्र 2 में दर्शाया गया है, हैप्लोटाइप को क्लस्टर माना जाता है। ए, बी, सी और डी मार्कोव श्रृंखला के विभिन्न सेटों को परिभाषित करते हैं जो हैप्लोटाइप वितरण



चित्र 2: क्रॉजक (2015) पर आधारित फास्टफेज के साथ जीनोटाइप इंप्यूटेशन।

उत्पन्न करते हैं। एलील रंग–कोडित होते हैं, और अलग–अलग रंग किसी दिए गए स्थान पर समान एलील के अनुरूप होते हैं, और साथ ही, FastPHASE में IMPUTE की तुलना में संभावना गणना में अधिक अज्ञात पैरामीटर शामिल हैं, साथ ही क्लस्टर वजन के प्रभाव, FastPHASE में पैरामीटर अनुमानों पर अधिक मानक त्रुटियां हैं। चरणबद्ध इम्प्युटशन आधार से प्राप्त मूल्यों पर कुछ मापदंडों को तय करना एक संभावित समाधान हो सकता है (मार्चिनी एंड होवी, 2010)।

मैच (Mach)

MaCH एक HMM मॉडल को नियोजित करता है जो कि IMPUTE के समान है, लेकिन, IMPUTE और FastPHASE के विपरीत, MaCH को एक अलग इंप्यूटेशन बेस की आवश्यकता नहीं है, जबकि इसमें वर्तमान अनुमानों के आधार पर प्रत्येक व्यक्ति के जीनोटाइप डेटा को चरणबद्ध अपडेट करके अन्य सभी नमूनों को अनुमानित किया जाता है। इस मॉडल को इस प्रकार लिखा जा सकता है:

$$P(G_i | D_{-i} \theta,\eta) = \sum P(G_i | Z,\eta)\ P(Z | D_{-i} \theta)$$

D_{-i}, को छोड़कर अनुमानित हैप्लोटाइप का सेट है, ZHMM की छिपी हुई अवस्था है, η यह निर्धारित करता है कि कॉपी किए गए हैप्लोटाइप से कितने समान हैं, और θ छिपी हुई अवस्था के बीच संक्रमण को नियंत्रित करता है (मार्चिनी और होवी, 2010)। समष्टि हैप्लोटाइप वितरण को पुनरावृत्त रूप से निर्धारित किया जाता है और प्रत्येक पुनरावृत्ति के दौरान पैरामीटर η और θ भी अपडेट किए जाते हैं। हैप्लोटाइप्स के एक संदर्भ पैनल H के आधार पर, अप्रबंधित जीनोटाइप को

इंप्यूटेशन अनुमानित हैप्लोटाइप्स D_{-i} में H को जोड़कर समायोजित किया जाता है। फिर हैप्लोटाइप नमूनों को पुनरावृत्त करके अप्रतिबंधित जीनोटाइप के सीमांत वितरण का अनुमान किया जाता है।

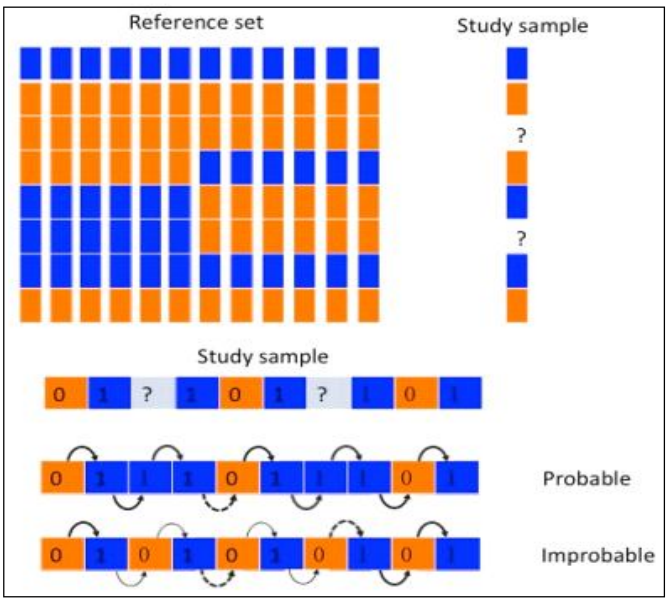
बीगल (BEAGLE)

बीगल विधि (ब्राउनिंग, 2006) एक गुणसूत्र के साथ प्रत्येक मार्कर की स्थिति में हैप्लोटाइप्स को स्थानीय रूप से समूह बनाकर एक HMM बनाती है। समूहों को एलील पर परिभाषित किया गया है न कि हैप्लोटाइप स्तर पर। FastPHASE के साथ एक और अंतर यह है कि समूहों की संख्या निश्चित नहीं होती बल्कि क्षेत्र–निर्भर है। इसके मॉडल में कोई पैरामीटर नहीं है जिसका अनुमान लगाने की आवश्यकता है और इसे दो चरणों में हैप्लोटाइप के दिए गए सेट पर लागू किया जाता है। वैश्विक हैप्लोटाइप आवृत्तियों के बजाय समीपस्थ मार्करों के बीच स्थानीय स्तर पर एलील एसोसिएशन संक्रमण की संभावनाओं को निर्धारित करता है। वंशानुक्रम पुनर्संयोजन प्रक्रिया को मॉडल करने की कोई आवश्यकता नहीं है क्योंकि मार्कोव श्रृंखला में कोई दीर्घकालिन स्मृति नहीं है, और किसी भी उत्परिवर्तन की अनुमति देने की आवश्यकता नहीं होती है। क्योंकि बीगल में मार्कोव श्रृंखला के एकमात्र अवस्था वास्तव में देखे गए एलील हैं। ये सभी विशेषताएं ऊपर बताए गए अन्य तरीकों की तुलना में जीनोटाइप इंप्यूटेशन में BEAGLE को कम्प्यूटेशनल रूप से तेज बनाती हैं। हालाँकि, दोष यह है कि BEAGLE को पूर्ण अध्ययन नमूने पर आधारित होना चाहिए, जिसका अर्थ है कि इसे IMPUTE2 की तरह विभाजित नहीं किया जा सकता है। यदि पड़ोसी मार्करों पर एक बेहिसाब एलील संयोजन वाले किसी भी हैप्लोटाइप को स्वचालित रूप से शून्य की संभावना सौंपी जाएगी।

बीगल विधि हैप्लोटाइप स्तर के बजाय एलील पर समूहों को परिभाषित करती है, एसएनपी के लिए, समूहों की संख्या प्रत्येक स्थिति (नीला या नारंगी) में दो के बराबर होती है। इंप्यूटेशन बेस और स्टडी सैंपल में इंटर–मार्कर एलील एसोसिएशन के दोनों स्थानीय स्तर हैप्लोटाइप वितरण संभावनाओं को निर्धारित करते हैं। टूटे हुए तीर मध्यवर्ती संक्रमण संभावनाओं को इंगित करते हैं (चित्र 3)।

इम्प्युटेशन विधियों के बीच तुलना

शोधकर्ताओं ने इन आमतौर पर इस्तेमाल किए जाने वाले जीनोटाइप इंप्यूटेशन फ्रेमवर्क (मार्चिनी एंड होवी, 2009; नोथनागेल और अन्य, 2009; वांग और अन्य, 2012; लियू और अन्य, 2014) के प्रदर्शन का सावधानीपूर्वक अध्ययन किया है। मार्चिनी एंड होवी (2010) ने संदर्भ पैनल के अनुसार सबसे लोकप्रिय इंप्यूटेशन विधियों में से प्रत्येक के गुणों को संक्षेप में प्रस्तुत किया है जो (तालिका 1), अध्ययन नमूनों के गुण (तालिका 2), कार्यक्रम विकल्प (तालिका 3), कम्प्यूटेशनल प्रदर्शन और त्रुटि दर (तालिका 4)। उन्होंने पाया कि IMPUTEv1, MACH, FastPHASE और BEAGLE के बीच IMPUTEv2 उनके सिमुलेशन परिदृश्य और एक MAR धारणा के आधार पर सबसे सटीक दृष्टिकोण है, लेकिन सभी विधियाँ समान प्रदर्शन उत्पन्न



चित्र 3: बीगल के साथ जीनोटाइप इंप्यूटेशन क्रॉजक (2015) पर आधारित FastPHASE से अलग।

करती हैं। इसके अलावा, आगे की जांच करने के लिए कि 1000 जीनोम प्रोजेक्ट जैसे हैप्लोटाइप्स के बड़े संदर्भ पैनल पर तरीके कैसे प्रदर्शन करते हैं, शोधकर्ताओं ने 1000 जीनोम प्रोजेक्ट से पायलट सीईयू हैप्लोटाइप के आधार पर 500 और 1,000 व्यक्तियों से युक्त 1,000 हैप्लोटाइप के संदर्भ पैनल का उपयोग किया है और HAPGEN सिमुलेशन लागू किया है। गुणसूत्र 10 पर 5 एमबी क्षेत्र। परिणाम दिखाते हैं कि IMPUTEv2, BEAGLE और FastPHASE की तुलना में तेज है।

वांग और अन्य, (2012) ने एक नई निकटतम पड़ोसी विधि (एनएन) और एक भारित संस्करण (डब्ल्यूएनएन) का प्रस्ताव रखा, जो दोनों सहसंयोजक सिद्धांत का पालन करते हैं। लक्षित व्यक्ति के पास जनसंख्या से एक के समान जीनोटाइप अनुक्रम होता है। इन दो विधियों के अलावा, उन्होंने FastPHASE, Npute (रॉबर्ट्स और अन्य, 2007) और कई मशीन लर्निंग इंप्यूटेशन विधियों को भी लागू किया, जिसमें सपोर्ट वेक्टर मशीन (SVM), एक स्थानीय न्यूरल नेटवर्क (न्यूरलनेट), और एक स्थानीय फर्स्ट ऑर्डर मार्कोव चेन (MC) शामिल हैं। परिणामों से पता चला कि एनएन और डब्ल्यूएनएन सबसे कुशल तरीकों में से थे, और लापता एसएनपी जीनोटाइप इंप्यूटेशन में फास्टफेस को छोड़कर अन्य तरीकों की तुलना में काफी बेहतर प्रदर्शन किया।

लियू और अन्य, (2014) 90 व्यक्तियों से पूरे–जीनोम डीएनए अनुक्रमण डेटा के आधार पर जीनोटाइप इंप्यूटेशन प्रदर्शन की व्यवस्थित रूप से जांच की। प्रतिरूपण प्रदर्शन का मूल्यांकन

तालिका 1: संदर्भ गुण।

गुण	इम्प्यूटेशन विधि				
	IMPUTEv1	IMPUTEv2-2	MACHv1-0-16	FastPHASEv1-4-0	BEAGLEv3-2
संदर्भ पैनल					
हैप्लोटाइप संदर्भ पैनल का उपयोग कर सकते हैं?	हाँ	हाँ	हाँ	हाँ	हाँ
जीनोटाइप्ड रेफरेंस पैनल का उपयोग कर सकते हैं?	नहीं	हाँ	हाँ	हाँ	हाँ
क्या दो हैप्लोटाइप या जीनोटाइप्ड रेफरेंस पैनल एक ही रन में इस्तेमाल किए जा सकते हैं?	नहीं	हाँ	नहीं	नहीं	नहीं
संदर्भ पैनल सही प्रारूप में उपलब्ध हैं	HapMap2 HapMap3 IKGP pilot data	HapMap2 HapMap3 IKGP pilot data 1000 Genome projects	HapMap2 HapMap3 IKGP pilot data 1000 Genome projects	Hap2	1000 Genome projects

तालिका 2: अध्ययन के नमूने के गुण।

गुण	इम्प्यूटेशन विधि				
	IMPUTEv1	IMPUTEv2-2	MACHv1-0-16	FastPHASEv1-4-0	BEAGLEv3-2
अध्ययन के नमूने					
अनिश्चितता के साथ निर्दिष्ट जीनोटाइप ले सकते हैं?	नहीं	हाँ	नहीं	नहीं	हाँ
तिकड़ी और संबंधित नमूनों को समायोजित कर सकते हैं?	नहीं	नहीं	नहीं	नहीं	हाँ
क्या ऑटोसोमल हैप्लोटाइप्स का एक अध्ययन नमूना लगाया जा सकता है?	हाँ	हाँ	नहीं	नहीं	हाँ
X गुणसूत्र पर इम्प्यूट कर सकते हैं?	हाँ	हाँ	नहीं	नहीं	हाँ

करने के लिए मानदंड के रूप में सटीक रूप से लगाए गए वेरिएंट के प्रतिशत का उपयोग करके, सबसे पहले, उन्होंने पाया कि मिनिमैक और IMPUTE2 में BEAGLE की तुलना में बेहतर इंप्यूटेशन प्रदर्शन है और बहु–जनसंख्या संदर्भ पैनल ने केवल संदर्भ पैनल का उपयोग करने की तुलना में बेहतर प्रदर्शन दिखाया है। एक ही आबादी। दूसरा, जांचकर्ता आमतौर पर आगे के विश्लेषण से खराब रूप से लगाए गए वेरिएंट को हटाने के लिए इंप्यूटेशन क्वालिटी माप पर भरोसा करते हैं।

तालिका 3: कार्यक्रम विकल्पों और सुविधाओं के गुण।

गुण	इम्प्यूटेशन विधि				
	IMPUTEv1	IMPUTEv2-2	MACHv1-0-16	FastPHASEv1-4-0	BEAGLEv3-2
प्रोग्राम के विकल्प और विशेषताएं					
क्या फेसिंग और साथ ही इम्प्यूटेशन उपलब्ध है ?	नहीं	हाँ	हाँ	हाँ	हाँ
क्या आंतरिक प्रदर्शन मूल्यांकन है?	हाँ	हाँ	हाँ	नहीं	नहीं
केवल एक निर्दिष्ट अंतराल में ही इम्प्यूट कर सकते हैं?	हाँ	हाँ	नहीं	नहीं	नहीं
डेटा सेट के बीच स्ट्रैंड संरेखण को संभाल सकता है?	हाँ	हाँ	हाँ	नहीं	नहीं
एसएनपी और नमूना समावेशन एवं बहिष्करण विकल्प?	हाँ	हाँ	नहीं	हाँ	हाँ
इंप्यूटेशन और एसोसिएशन परीक्षण के लिए संयुक्त मॉडल?	नहीं	नहीं	नहीं	नहीं	नहीं
ऑपरेटिंग सिस्टम की आवश्यकता?	Linux, Solaris, Windows, Mac	Linux, Solaris, Windows, Mac	Linux, Windows, Mac	Linux, Solaris, Windows, Mac	Java executable

तालिका 4: कम्प्यूटेशनल प्रदर्शन।

गुण	इम्प्यूटेशन विधि				
	IMPUTEv1	IMPUTEv2-2	MACHv1-0-16	FastPHASEv1-4-0	BEAGLEv3-2
कम्प्यूटेशनल प्रदर्शन					
आकलन 1*	नहीं	हाँ	नहीं	नहीं	हाँ
आकलन 2*	नहीं	नहीं	नहीं	नहीं	हाँ
त्रुटि दर¢					
पंक्तियाँ परिदृश्य A के अनुरूप हैं	5.42%	5.16%	5.46%	5.92%	6.33%
परिदृश्य बी (प्रतिबंधित)	....	3.40%	....	5.33%	3.46%
परिदृश्य बी (पूर्ण)	....	3.40%	....		4.01%

★120 CEU Hapmap2 haplotypes से Affy500k चिप पर 1377 नमूनों का इम्प्यूटेशनय 7.5 एमबी क्षेत्र। डेटा (होवी और अन्य, 2009) से आता है। रु 500 (1000) नमूनों का इम्प्यूटेशन 5 एमबी क्षेत्र में 8712 एसएनपी पर 872 एसएनपी से 872 एसएनपी पर जीनोटाइप किया गया। क्रोमोसोम 10 पर 5 एमबी क्षेत्र में 1000 जीनोम परियोजना से HAPGEN और पायलट CEU हैप्लोटाइप का उपयोग करके सिम्युलेटेड डेटासेट पर आधारित समय। ¢ त्रुटि दर (होवी और अन्य, 2009), IMPUTEv2 के परिणाम अपडेट कर दिए गए हैं। दिए गए परिदृश्य B त्रुटि दर Affymetrix SNPs से लगाए गए Illumina SNPs के लिए हैं। Illumina एसएनपी से लगाए गए एफिमेट्रिक्स एसएनपी के लिए त्रुटि दर ब्रैकेट में दी गई है।

निष्कर्ष

इस अध्ययन में विचार किए गए मिसिंग डेटा परिदृश्यों में IMPUTEv2 में उच्च सटीकता और तेज गणना दोनों थे। होवी और अन्य, (2011) का मानना है कि IMPUTE2 की सफलता इसकी कम्प्यूटेशनल रणनीतियों और अपने डीएनए अनुक्रम भिन्नता के मॉडल की वजह से है। मॉडल के एल्गोरिथ्म के दृष्टिकोण से, BEAGLE और FastPHASE हैप्लोटाइप को समूहों में जोड़ते हैं जो गणना प्रक्रिया को गति देते हैं क्योंकि यह HMM अवस्थाओं की संख्या को सीमित करता है। BEAGLE और FastPHASE को डेटासेट में प्रत्येक हैप्लोटाइप पर HMM गणना करने के बजाय केवल समूहों के एक छोटे सेट पर गणना चलाने की आवश्यकता होती है। इसके विपरीत, IMPUTEv2 और MaCH एक स्टेट में प्रत्येक हैप्लोटाइप के लिए HMM का प्रदर्शन करते है: हालांकि, सभी स्टेट का उपयोग गणना को कठिन बना देता है और इसलिए IMPUTEv2 राज्यों अवस्थाओं को प्रतिबंधित करता है।

होवी और अन्य, (2011) ने प्रदर्शित किया कि IMPUTEv2 उनके द्वारा जांचे गए परिदृश्यों के लिए BEAGLE की तुलना में उच्च सटीकता प्राप्त करता है, और यह डेटासेट में कम–आवृत्ति वाले वेरिएंट में विशेष रूप से स्पष्ट है जिसमें उच्च हैप्लोटाइप विविधता है। जीनोटाइप इंप्यूटेशन जिसने जटिल मानव रोगों को सैंकड़ों वास्तविक संघों तक पहुँचाया है, जीनोम–वाइड एसोसिएशन स्टडीज (GWAS) का एक आवश्यक हिस्सा बन रहा है (हिंडोर्फ और अन्य, 2009) और यह अगले कुछ वर्षों तक जारी रहेगा। मुख्य कारक जो प्रभावित करेगा कि किस इंप्यूटेशन विधि का उपयोग किया जाता है, वे होंगे जो अगली पीढ़ी के अनुक्रमण डेटा की बढ़ती उपलब्धता को बड़ी संख्या में एसएनपी के साथ संभाल सकते हैं। हैप्लोटाइप के बड़े, अधिक विविध सेट का उपयोग करने में आरोपण विधियों के लिए भी यही चुनौती है।

अंत में, जीनोटाइप इंप्यूटेशन को प्रयोगशाला–आधारित डेटा पीढ़ी जैसे अच्छे वैज्ञानिक अभ्यास के समान नियमों का पालन करना चाहिए। आनुवंशिक रोग संघ अध्ययन (एंडरसन और अन्य, 2010) में डेटा गुणवत्ता के मानदंड को परिभाषित करने के लिए अतीत में बहुत सारे प्रयास किए गए हैं। इसलिए, जीनोटाइप इंप्यूटेशन के लिए समान स्तर की सटीकता और विश्वसनीयता मूल्यांकन मानदंड विकसित करना आवश्यक है।

संदर्भ

Anderson, C.A., Pettersson, F.H. and Clarke, G.M. (2010). Data quality control in genetic case-control association studies. Nature Protocols. 5: 1564-1573.

Browning, S.R. (2006). Multilocus association mapping using variable-length Markov chains. American Journal of Human Genetics. 78: 273-280.

Cheema, J.R. (2014). A review of missing data handling methods in educational research. Review of Educational Research. XX(X): 1-22.

Fearnhead, P. and Donnelly, P. (2001). Estimating recombination rates from population genetic data. Genetics. 159: 1299-1318.

Hindorff, L.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences. 106(23): 9362-9367.

Howie, B. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature Genetics. 44(8): 955-960.

Howie, B., Marchini, J. and Stephens (2011). Genotype imputation with thousands of genomes. G3: Genes, Genomes, Genetics. 1: 457-469.

Howie, B.N., Donnelly, P. and Marchini, J. (2009). A flexible and accurate method for the next generation of genome-wide association studies. Plos Genetics. 5(6): e1000529. doi:10.1371/journal.pgen.1000529.

Krawczak, M. (2015). Genotype Imputation. In: eLS. John Wiley and Sons, Ltd: Chichester. doi: 10.1002/9780470015902.a0022399.

Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. 165: 2213-2233.

Liu, Q. (2014). Systematic assessment of imputation performance using the 1000 Genomes reference panels. Briefings in Bioinformatics. DOI: 10.1093/bib/bbu035.

Marchini, J. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics. 39(7): 906-913.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. Nature Reviews Genetics. 11: 499-511.

Nothnagel, M. (2009). A comprehensive evaluation of SNP genotype imputation. Human Genetics. 125: 163-171.

Roberts, A. (2007). Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. Bioinformatics. 23: 401-407.

Roth, P. (1994). Missing data: A conceptual review for applied psychologists. Personnel Psychology. 47: 537-560. doi: 10.1111/j.1744-6570.1994.tb01738.x.

Rubin, D.B. (1976). Inference and missing data. Biometrika. 63(3): 581-592.

Servin, B. and Stephens, M. (2007). Imputation-based analysis of association studies: Candidate Regions and Qualitative Traits. Plos Genetics. 3(7): e114. doi: 10.1371/journal.pgen. 0030114.

Sheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. The American Journal of Human Genetics. 78: 629-644.

Wang, Y.N., Cai, Z.P., Stothard, P. (2012). Fast accurate missing SNP genotype local imputation. BMC Research Notes. 5: 404. doi: 10.1186/1756-0500-5-404.