SNP-BLUP, G-BLUP ja H-BLUP - johdanto genomisiin arvosteluihin

Minna Koivula, Esa Mäntysaari ja Ismo Strandén

MTT, Biotekniikka- ja elintarviketutkimus, Biometrinen genetiikka, 31600 Jokioinen
e-mail. etunimi.sukunimi@mtt.fi

Tiivistelmä

Genominen valinta tarkoittaa eläimen (ja kasvien) jalostusarvon ennustamista ja valintapäätöksen tekemistä yksilön DNA:n eli SNP- markkeritietojen perusteella. Viime vuosina genomisten jalostusarvojen laskentaan on kehitelty useita malleja. Yksinkertaisessa DNA markkeri eli SNP-BLUP menetelmään perustuvassa mallissa SNP-markkerivaikutuksia käsitellään satunnaisina ja markkerien geneettisen varianssin oletetaan olevan sama kaikille markkereille. Tällöin yksilölle arvioitu genominen jalostusarvo saadaan laskemalla yhteen kaikki sen markkerivaikutukset. Genomisen sukulaisuusmatriisin (**G**) käyttö jalostusarvosteluissa mahdollistaa genomisen informaation hyödyntämisen perinteisen kaltaisessa jalostusarvostelumallissa, ns. G-BLUP:ssa. Tällöin genotyypitettyjen eläinten väliset sukulaisuudet saadaan tarkemmin kuin tavallisessa sukupuussa: esim. täyssisarten kesken genominen sukulaisuusaste voi vaihdella. Genomiset jalostusarvot SNP-BLUP mallilla ja **G**-matriisiin perustuvalla G-BLUP mallilla ovat samoja. Genomisten jalostusarvojen arvosteluvarmuutta voidaan parantaa yhdistämällä genominen informaatio perinteiseen jalostusarvoon. Tämä voidaan tehdä joko käsittelemällä genomiset jalostusarvot eri ominaisuuksina tai yhdistämällä genomitiedot suoraan perinteisiin jalostusarvosteluihin. Jälkimmäisessä ns. single-step - menetelmässä (H-BLUP) genotyypitettyjen eläinten sukulaisuudet perustuvat genotyyppi-tietoihin ja muiden eläinten sukulaisuudet sukupuutietoihin. Testasimme eri genomisia malleja pohjoismaisella punaisella rodulla. Tulosten mukaan eri genomiset arvostelumallit antavat vertailukelpoisia tuloksia, joten kulloinkin käytettävä menetelmä voi perustua käytännön seikkoihin. SNP- BLUP ja G-BLUP antoivat samat genomiset jalostusarvot, mutta ne erosivat hieman H-BLUP:n antamista arvoista. Keskimäärin genomisten jalostusarvojen luotettavuus oli 12 prosenttiyksikköä korkeampi kuin perinteisistä jalostusarvoista lasketun polveutumisindeksin. Lisäksi H-BLUP:lla genotyypitettyjen sonnien arvosteluvarmuus oli 2 - 4 prosenttiyksikköä korkeampi kuin muilla genomisilla malleilla.

Asiasanat: Genominen valinta, SNP, genomiset arvostelut, **G**-matriisi

Johdanto

Perimän kartoittamiseen tarkoitetut laboratoriomenetelmät ovat kehittyneet vauhdilla viime vuosina. Nykyisin on mahdollista määrittää eläimeltä useita tuhansia geenimerkkejä nopeasti ja taloudellisesti. Tämä on mahdollistanut ns. genomisen valinnan käytön jalostuksessa. Genominen valinta tarkoittaa eläimen (ja kasvien) jalostusarvon ennustamista ja valintapäätöksen tekemistä yksilön koko genomin tai oikeastaan siitä määriteltyjen geenimerkkien perusteella. Käytetyt geenimerkit ovat ns. SNP:tä (single-nucleotide polymorphism), jotka tunnistavat DNA-ketjun yksittäisessä emäksessä esiintyvän vaihtelun. DNA-markkerien käyttö yhtenä eläinten valintakriteerinä onkin muodostunut merkittäväksi jalostuksen työvälineeksi. Periaatteena on, että DNA-testistä saatu genomitieto yhdistetään perinteiseen, sukupuuta hyödyntävään ominaisuuksien mittaustietoon perustuvaan jalostusarvosteluun. Genomiset jalostusarvot voidaan laskea eläimelle heti sen synnyttyä. Näin ollen suurin hyöty genomisesta arvostelusta saadaan nuorille eläimille, joilta ei ole vielä omia tuloksia tai jälkeläisiä. Tämä tehostaa erityisesti alhaisen periytymisasteen ominaisuuksien, kuten terveys- ja lisääntymisominaisuuksien jalostusarvostelua.

Tällä hetkellä useimmat käytössä olevat genomiset arvostelumallit ovat kaksi- tai kolmivaiheisia (VanRaden 2008, Hayes ym. 2009, VanRaden ym. 2009). Ensin eläimille lasketaan perinteiset jalostusarvot, sitten yhdistetään jälkeläisten perusteella jalostusarvosteltujen sonnien jalostusindeksit ja SNP-markkeritiedot, jolloin saadaan genomisen jalostusarvon ennustemalli. Tämän jälkeen voidaan ennustaa jalostusarvot niille eläimille, joilta on genomitietoa, mutta joiden varsinainen jalostusindeksi perustuu ainoastaan vanhempien jalostusarvoihin.

Viime vuosina genomisten jalostusarvojen laskentaan on kehitelty useita malleja. Yksinkertaisessa markkeri eli SNP-BLUP menetelmään perustuvassa mallissa SNP-markkerivaikutuksia käsitellään satunnaisina ja varianssin oletetaan olevan sama kaikille markkereille (Meuwissen ym. 2001, VanRaden 2008). Oletuksen mukaan siis jokaisella SNP-merkillä voi olla yhtä suuri vaikutus eläimen jalostusarvoon. Eläimen genominen jalostusarvo voidaan tämän jälkeen laskea summaamalla yhteen kaikki sen markkerivaikutukset. Genomista sukulaisuusmatriisia (**G**) käyttämällä genomista informaatiota voidaan hyödyntää perinteisen kaltaisessa jalostusarvostelumallissa, ns. G-BLUP:ssa (Goddard 2009). Tällöin genotyypitettyjen eläinten väliset sukulaisuudet saadaan tarkemmin kuin tavallisessa sukupuussa: esim. täyssisarten kesken genominen sukulaisuusaste voi vaihdella. On osoitettu (Strandén ja Garrick 2009), että em. oletuksilla SNP-BLUP menetelmä ja G-BLUP menetelmä ovat yhdenmukaisia. Genomisten jalostusarvojen arvosteluvarmuutta voidaan edelleen parantaa yhdistämällä genominen informaatio perinteiseen jalostusarvoon. Tämä voidaan tehdä joko käsittelemällä genomiset jalostusarvot eri ominaisuuksina tai yhdistämällä genominen markkeritieto suoraan perinteisiin jalostusarvosteluihin. Jälkimmäisessä ns. single-step - menetelmässä (H-BLUP) genotyypitettyjen eläinten sukulaisuudet perustuvat genotyyppi-tietoihin ja muiden eläinten sukulaisuudet sukupuutietoihin (Aguilar ym. 2010, Christensen ja Lund 2010).

Tämän tutkimuksen tarkoituksena oli verrata eri genomisia malleja pohjoismaisella punaisella rodulla.

Aineisto ja menetelmät

Aineisto

Pohjoismaisen punaisen rodun sonnit genotyypitettiin käyttämällä Illumina Bovine SNP50 BeadChip-lastua (Illumina, San Diego, CA). Puuttuvat markkerit lisättiin imputoimalla fastPHASE-ohjelmalla (Scheet ja Stephens 2006). Lopullisessa genomiaineistossa oli 6145 genotyypitettyä pohjoismaista punaisen rodun sonnia ja näillä kaikilla 37996 SNP-markkeria.

Eläinten fenotyyppiset havainnot saatiin maaliskuun 2010 pohjoismaisista jalostusarvosteluista (NAV) ja aineistossa oli sonnien jalostusarvot (EBV) tuotanto-ominaisuuksille sekä utareterveydelle, arvosteluvarmuudet (r^2_{EBV}), sekä sonnien tytärmäärät (EDC). Eri indeksien luotettavuus arvioitiin ns. validointimenetelmällä, jossa arvioidaan markkerivaikutukset vanhempia sonneja sisältävällä referenssijoukolla ja ennustetaan nuorempia kandidaattisonneja. Tällaisessa katkaistussa aineistossa samojen sonnien EBV:t oli laskettu käyttäen tietoja vuoteen 2005 saakka, eli

viiden vuoden havainnot oli poistettu. Sonnit joiden EBV oli 2010 aineistossa laskettu vähintään 10 tyttären perusteella ja joilla katkaistussa 2005 aineistossa oli vain polveutumisindeksi, luokiteltiin kandidaattisonneiksi. Analyyseissä käytettiin vastemuuttujina deregressoituja jalostusarvoja (DRP). DRP:t laskettiin DeRegress- optiolla (Strandén ja Mäntysaari 2010) MiX99 ohjelmassa (Strandén ja Lidauer 1999). Deregressoinnissa sukupuun vaikutus EBV-havainnoista poistetaan ja samalla palautetaan sonnien tytärryhmien koosta aiheutuva vaihtelu takaisin vastemuuttujiin. Sonnien määrät eri luokissa annetaan taulukossa 1.

Taulukko 1. Sonnit (n), joilla DRP eri osa-aineistoissa, ominaisuuksittain, sekä deregressiossa käytetyt periytymisasteet

	2010 Data		2005 Data			
			Referenssisonnit		Kandidaatti-sonnit	Periytymisaste
	Genotyypitetyt		Genotyypitetyt			h^2
Maito	6253	4145	5313	3330	809	0.39
Valkuainen	6253	4145	5313	3330	809	0.31
Rasva	6253	4145	5313	3330	809	0.36
Utareterveys	6169	4431	5363	3649	780	0.04

Tilastolliset menetelmät

Genomisilla malleilla arvioitiin aluksi jokaisen SNP-markkerin vaikutus referenssipopulaation avulla. Kun näillä ratkaisuilla kerrotaan valintakandidaattien SNP markkerigenotyypien arvot, saadaan suorat genomiset arvot (DGV) kandidaattisonneille.

Laskennassa käytettiin MiX99-ohjelmistoa (Strandén ja Lidauer 1999). Kaikissa malleissa käytettiin DRP:lle painokertoimena EDC_i:tä, joka on skaalattu EDC. Ensimmäinen malli on yksinkertainen SNP markkerimalli (SNP-BLUP).

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Mg} + \mathbf{e} \qquad (1)$$

missä **y** on genotyypitettyjen sonnien DRP-vektori, μ on keskiarvo, **g** on satunnaisten markkerivaikutusten vektori ja **e** on jäännösvektori. SNP-BLUP markkeriratkaisuista voidaan laskea DGV:t: $\hat{\mathbf{a}} = \mathbf{1}\hat{\mu} + \mathbf{M}\hat{\mathbf{g}}$.

G-BLUP vastaa menetelmänä SNP-BLUP:a, mutta markkerivaikutusten sijaan malliin sijoitetaan DGV-vektori jonka kovarianssirakenteena on genominen sukulaisuusmatriisi **G** sukulaisuusmatriisin polveutumiseen perustuvan **A**:n sijasta:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Xa} + \mathbf{e} \qquad (2)$$

missä **a** on DGV-vektori ja **X** on DGV havaintoja vastaava insidenssimatriisi. Malli on siis vastaava kuin tavallinen eläinmalli mutta käyttää genomista sukulaisuusinformaatiota.

Single-step arvosteluissa on genotyypitetyille eläimille ja niiden sukulaisille kaikille oletettu samankaltainen genominen jalostusarvo. Erotuksena G-BLUP mallista, näillä jalostusarvoilla on erilainen varianssirakenne (H-sukulaisuusmatriisi) riippuen siitä onko eläin genotyypitetty vai ei. Single-step H-BLUP tuottaa DGV:n sijasta GEBV:n (genominen jalostusarvoindeksi).

$$\mathbf{y}_t = \mathbf{1}\mu + \mathbf{Wa} + \mathbf{e} \qquad (3)$$

missä $\mathbf{y}_t$ on kaikkien sonnien DRP vektori, **a** on additiivisten geneettisten tekijöiden vektori ja **W** on mallimatriisi, joka yhdistää havainnot ja genomiset arvot.

Mallien vertailu perustui Interbull GEBV validointitestiin (Mäntysaari ym. 2010). Validoinnissa arvioidaan genomisten arvostelujen keskimääräisen arvosteluvarmuus. Ensin lasketaan genomiset ennusteet katkaistulla aineistolla. Tämän jälkeen deregressoidaan koko aineistosta (2010 aineisto) kandidaattisonnien jalostusarvot ja regressoidaan nämä genomisten arvostelujen ennusteisiin. Odotusarvo tälle regressiolle on 1.0, eli että sonnien väliset erot genomisissa indekseissä näkyvät yhtä suurina niiden tyttärien tuotoskeskiarvoissa.

Tulokset ja tulosten tarkastelu

SNP-BLUP mallilla ja **G**-matriisiin perustuvalla G-BLUP mallilla saadut ennusteet ovat hyvin samankaltaiset. Korrelaatiot mallien antamien ennusteiden välillä olivat 0.999. H-BLUP:lla saadut genomiset ennusteet erosivat hivenen muiden mallien antamista ennusteista ja korrelaatio H-BLUP:n ja muiden arvostelumallien välillä oli ominaisuudesta riippuen 0.979 – 0.998. Tämä on odotettavissa, koska H-BLUP:ssa käytetään sukupuutietoja ja havaintoja sekä genotyypitetyiltä että genotyypittämättömiltä eläimiltä. Koska mallit antavat vertailukelpoisia tuloksia, menetelmän valinta voi perustua käytännön seikkoihin. SNP-BLUP on yksinkertainen tapa saada eläimille genomiset jalostusarvot heti, kun eläimelle on olemassa markkeritiedot. G-BLUP:lla puolestaan on helpompi laskea eläinkohtainen teoreettinen arvosteluvarmuus, mutta **G**-matriisin rakentaminen ja kääntäminen on laskennallisesti vaativaa referenssipopulaation ollessa suuri.

Taulukossa 2 esitetään eri malleille regressionmallin vakiotermi (b_0), regressiokerroin (b_1) sekä arvosteluvarmuudet (R^2). Jotta saadut ennusteet olisivat harhattomia, regressiomallin vakion (b_0) pitäisi olla lähellä EBV:n keskiarvoa (tässä tapauksessa 0) ja regressiokertoimen (b_1) pitäisi olla lähellä yhtä. Taulukosta 2 kuitenkin voi havaita, että SNP-BLUP:lla ja G-BLUP:lla regressiokertoimet olivat 0.76 - 0.78 maitotuotoksella, valkuaistuotoksella ja utareterveydellä, sekä 0.85 rasvatuotoksella. Alhaiset regressiokertoimet antavat olettaa, että SNP-BLUP:n ja G-BLUP:n antamat DGV:t yliarvioivat jalostusarvojen välistä vaihtelua. Toisaalta myös polveutumisindeksin (PA) antamat b_0 ja b_1 luvut osoittavat, että nuorten sonnien saamat polveutumisindeksit ovat vähän odotettua suurempia, ja vaihtelu on suurempaa. Single-step H-BLUP:lla regressiokertoimet ovat hivenen lähempänä yhtä, mikä tukee ajatusta, että H-BLUP parantaa genomisten arvostelujen luotettavuutta.

Arvosteluvarmuudet SNP-BLUP:lla ja G-BLUP:lla olivat 0.30 - 0.31 maidolle ja valkuaiselle, mutta hivenen korkeampia rasvalle (0.40). Tämä on oletettavasti rasvaprosenttiin vaikuttavan DGAT geenin seurausta (Grisart 2004). Utareterveydellä arvosteluvarmuudet olivat huomattavasti alhaisempia kuin tuotosominaisuuksille. Tämä voi osittain johtua utareterveyden alhaisesta periytymisasteesta ja siitä, että DRP:n luotettavuus utareterveydellä on myös alhaisempi kuin tuotosominaisuuksilla. Kaiken kaikkiaan genomiset arvostelut kuitenkin nostivat arvosteluvarmuutta huomattavasti verrattuna polveutumisindeksiin. Maidolla ja valkuaisella nousu oli keskimäärin 11 prosenttiyksikköä, utareterveydellä 9 prosenttiyksikköä ja rasvalla 17 prosenttiyksikköä.

Single-step menetelmällä genotyypitettyjen sonnien arvosteluvarmuus oli tuotanto-ominaisuuksilla noin 2 – 4 prosenttiyksikköä korkeampi verrattuna muihin genomiarvosteluihin. Utareterveyden arvostelussa single-step ei kuitenkaan tuonut mitään hyötyä genotyypitettyjen eläinten arvosteluvarmuuteen. Single-step H-BLUP näyttää kuitenkin varteenotettavalta vaihtoehdolta genomisten arvostelujen laskemiseen.

Taulukko 2. Regressiomallin vakiotermi (b_0), regressiokerroin (b1) sekä arvosteluvarmuudet (R^2) SNP-BLUP, G-BLUP ja H-BLUP malleista ja polveutumisindeksistä (PA) eri ominaisuuksille.

	Maito			Valkuainen			Rasva			Utareterveys		
	b_0	b_1	R^2	b_0	b_1	R^2	b_0	b_1	R^2	b_0	b_1	R^2
PA	3.28	0.73	0.19	4.26	0.77	0.20	2.34	0.83	0.23	0.05	0.66	0.08
SNP-BLUP	3.17	0.76	0.30	4.49	0.77	0.31	2.25	0.85	0.40	0.54	0.76	0.17
G-BLUP	3.10	0.77	0.30	4.54	0.78	0.31	2.20	0.86	0.40	0.57	0.77	0.17
H-BLUP	2.73	0.80	0.32	3.63	0.83	0.34	1.93	0.90	0.42	0.38	0.78	0.17

Johtopäätökset

Tulosten mukaan eri genomiset arvostelumallit antavat vertailukelpoisia tuloksia, joten kulloinkin valittava menetelmä voi perustua käytännön seikkoihin. SNP- BLUP ja G-BLUP antoivat samat genomiset jalostusarvot, mutta ne erosivat hieman H-BLUP:n antamista arvoista. Keskimäärin genomisten jalostusarvojen luotettavuus oli 12 prosenttiyksikköä korkeampi kuin perinteisellä polveutumisindeksillä. Lisäksi H-BLUP:lla genotyypitettyjen sonnien arvosteluvarmuus oli 2 - 4 prosenttiyksikköä korkeampi kuin muilla genomisilla malleilla. Single-step menetelmästä saatavan

hyödyn ei voida olettaa olevan kovin suuri käytettäessä arvosteluissa pelkkiä sonnien jalostusarvoja. Tällöin sonnin polveutumisindeksi PA perustuu vain sen isän, ja emänisän tuloksiin, ja nämä saattavat olla myös genotyypitettyjä. Hyödyllisempää on jos single-step arvosteluissa käytetään suoraan eläinmalli sukupuuta, jolloin myös genotyypittämättömien sonninemien tulokset yhdistyvät GEBV–indekseihin (Mäntysaari ym. 2011). Vielä paremmin informaatiota hyödynnetään, jos single-step arvostelussa käytetään suoraan alkuperäisiä havaintoja deregressoitujen jalostusarvojen sijasta. Tällöin genominen informaatio tulee huomioiduksi myös ympäristövaikutuksia ratkaistaessa.

Kirjallisuus

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A. & Tsuruta, S. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93:743-752.

Christensen, O.F. & Lund, M.S. 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42:2.

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136:245-257.

Grisart, B., Farnir, F.. Karim, L., Cambisano, N. & Kim, J.J. 2004. Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. Proc. Natl. Acad. Sci. USA 101: 2398-2403

Hayes, B.J., Bowman, P.J, Chamberlain, A.J. & Goddard, M.E. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92: 433-443.

Meuwissen, T.H.E., Hayes, B.J. & Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-1829.

Mäntysaari, E.A., Liu, Z. & VanRaden, P. 2010. Interbull validation test for genomic evaluations. Interbull Bull. 40: 1-5.

Mäntysaari, E.A.,Koivula, M., Pösö, J., Aamand, G.P. & Strandén, I. 2011.Estimation of GEBVs using deregressed individual cow breeding values. Interbull open meeting 27.-28. August 2011, Stavanger, Norway.

Scheet, P. & Stephens, M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. 78: 629-644.

Strandén, I. & Garrick, D. 2009. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J. Dairy Sci. 92: 2971-2975.

Strandén, I. & Lidauer, M. 1999. Solving large mixed models using preconditioned conjugate gradient iteration. J. Dairy Sci. 82: 2779-2787.

Strandén, I. & Mäntysaari, E.A. 2010. A recipe for multiple trait deregression. Interbull Bull. 42: 21-24.

VanRaden, P.M. 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91: 4414-4423.

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S. & Schnabel, R.D. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92: 16-24.