Supplementary Materials: Multitask group Lasso for Genome Wide association Studies in diverse populations

Asma Nouira* and Chloé-Agathe Azencott
MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, F-75006 Paris, France
Institut Curie, PSL Research University, F-75005 Paris, France
INSERM, U900, F-75005 Paris, France
*E-mail: asma.nouira@mines-paristech.fr**

Appendix A. Data availability

Appendix A.1. *Simulated data*

Code to reproduce our simulations is available on `https://github.com/asmanouira/MuGLasso_GWAS`

Table A1 shows the location of the predefined disease loci, for each population. Table A2 shows the number of predefined disease loci, both common to both population and specific to each population.

Table A1. For simulated data, location of predefined disease loci represented by start/end positions information in each subpopulation through chromosomes: 2, 12, 19, 21 and 22.

Chromosome	Subpopulations	
	CEU	YRI
2	1 000 - 50 000	1 000 - 50 000
12	10 - 37 000	10 - 40 000
19	1 000 - 50 000	1 000 - 50 000
21	10 - 10 000	10 - 7 000
22	-	10 - 2 000

Table A2. For simulated data, number of predefined causal SNPs

Populations	Number of SNPs
Specific-CEU	2 999
Specific-YRI	4 989
Shared (CEU+YRI)	141 982
Total	149 970

Appendix A.2. *DRIVE*

Data access The dataset “General Research Use” in DRIVE Breast Cancer OncoArray Genotypes is available from the dbGaP controlled-access portal, under Study Accession phs001265.v1.p1 (`https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\_id=phs001265.v1.p1`). Researchers can gain access the data by applying to the data access committee, see `https://dbgap.ncbi.nlm.nih.gov`.

Ethics approval The dataset was obtained from NIH after ethical review of project #17707, titled ”Network-guided multi-locus biomarker discovery”, and used under approval of this request (#67806-4).

Acknowledgments OncoArray genotyping and phenotype data harmonization for the Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) breast-cancer case control samples was supported by X01 HG007491 and U19 CA148065 and by Cancer Research UK (C1287/A16563). Genotyping was conducted by the Center for Inherited Disease Research (CIDR), Centre for Cancer Genetic Epidemiology, University of Cambridge, and theNational Cancer Institute. The following studies contributed germline DNA from breast cancer cases and controls: the Two Sister Study (2SISTER), Breast Oncology Galicia Network (BREOGAN), Copenhagen GeneralPopulation Study (CGPS), Cancer Prevention Study 2 (CPSII), The European Prospective Investigation intoCancer and Nutrition (EPIC), Melbourne Collaborative Cohort Study (MCCS), Multiethnic Cohort (MEC), NashvilleBreast Health Study (NBHS), Nurses Health Study (NHS), Nurses Health Study 2 (NHS2), Polish Breast CancerStudy (PBCS), Prostate Lung Colorectal and Ovarian Cancer Screening Trial (PLCO), Studies of Epidemiologyand Risk Factors in Cancer Heredity (SEARCH), The Sister Study (SISTER), Swedish Mammographic Cohort (SMC), Women of African Ancestry Breast Cancer Study (WAABCS), Women’s Health Initiative (WHI).

Appendix B. Supplementary Methods

Appendix B.1. *LD groups across populations*

Figure B1 illustrates the process by which we obtain LD-groups across populations, from LD-groups obtained on each population separately using adjacency-constrained hierarchical clustering (see Section 2.2.1)

Appendix B.2. *Multitask group lasso*

Figure B2 illustrates the architecture of the multitask group Lasso described in Section 2.3.

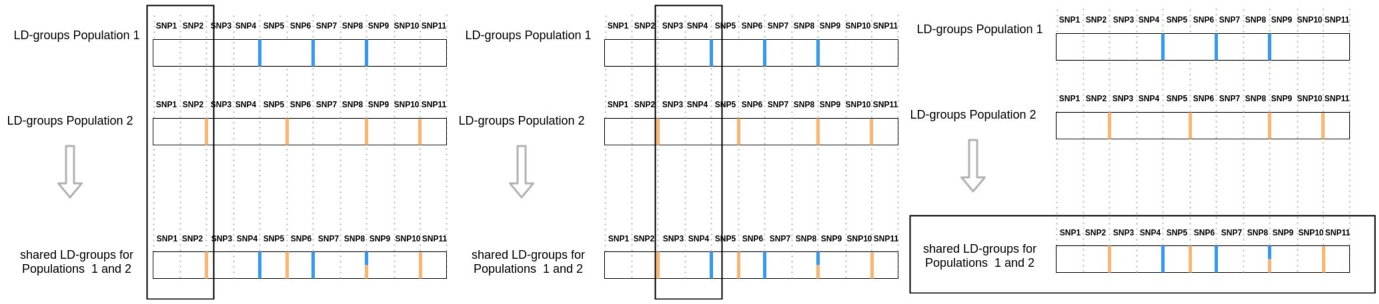

LD-groups Population 1
LD-groups Population 2
shared LD-groups for Populations 1 and 2
SNP1 SNP2 SNP3 SNP4 SNP5 SNP6 SNP7 SNP8 SNP9 SNP10 SNP11

Fig. B1. Choice of shared LD-groups choice after adjacency-constrained hierarchical clustering for each population

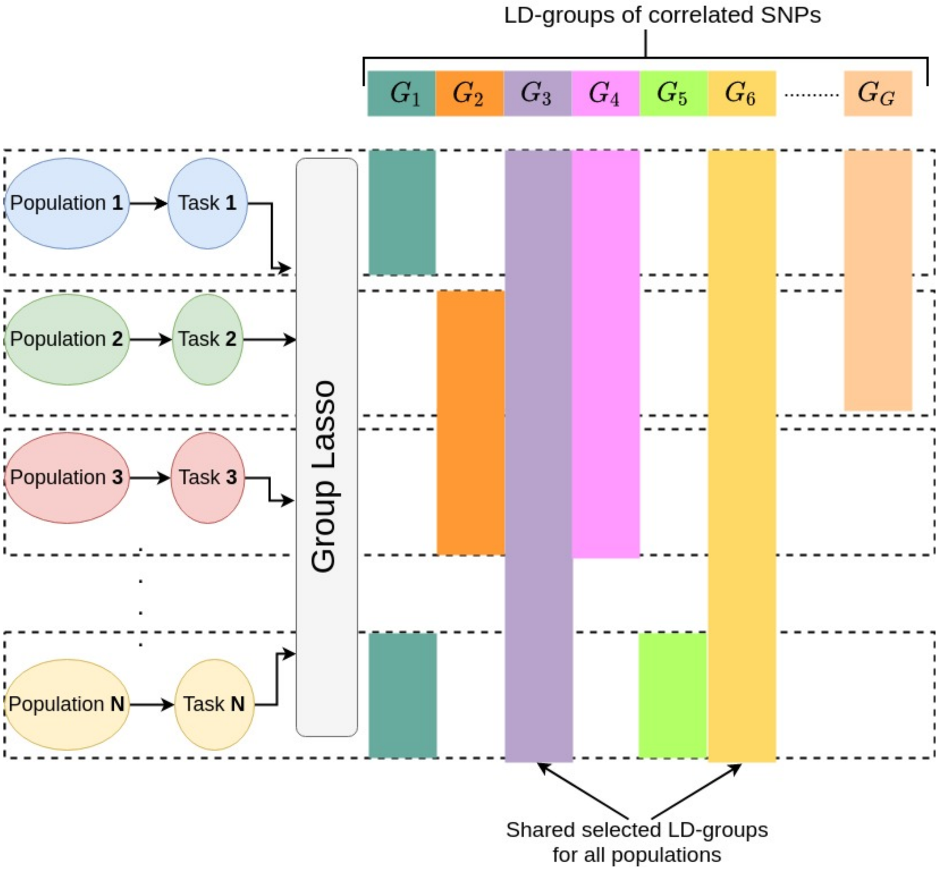

LD-groups of correlated SNPs
G_1 G_2 G_3 G_4 G_5 G_6 .... G_G
Population 1
Task 1
Population 2
Task 2
Population 3
Task 3
Population N
Task N
Group Lasso
Shared selected LD-groups for all populations

Fig. B2. Multitask group Lasso architecture

Appendix B.3. *Gap safe screening rules*

Let $X \in \mathbb{R}^{n\times d}$ be a design matrix and $\boldsymbol{y} \in \mathbb{R}^n$ the corresponding vector of outcomes, which can be binary or real-valued. We consider the following optimization problem:

$$\widehat{\boldsymbol{\beta}}^{(\lambda)} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\arg\min}\, P_\lambda(\boldsymbol{\beta}) := \sum_{i=1}^{n} f_i\left(X_{i.}^\top \boldsymbol{\beta}\right) + \lambda\,\Omega(\boldsymbol{\beta}), \tag{B.1}$$

where all $f_i : \mathbb{R} \to \mathbb{R}$ are convex and differentiable functions with $1/\gamma-$Lipschitz gradient, and $\Omega : \mathbb{R}^d \to \mathbb{R}_+$ is a norm that is group-decomposable, i.e., the set of d features is partitioned in G groups of sizes $d_1, d_2, \ldots, d_G$, and

$$\Omega(\boldsymbol{\beta}) = \sum_{g=1}^{G} \Omega_g \left(\boldsymbol{\beta}^{(g)} \right),$$

where each Ω_g is a norm on $\mathbb{R}^{d_g}$ and, as previously, $\boldsymbol{\beta}^{(g)}$ corresponds to the coefficients of $\boldsymbol{\beta}$ restricted to the features in group g. As before, the λ parameter is a non-negative constant controlling the trade-off between the data fitting term and the regularization term.

Equation (2) is a special case of Equation (B.1) because the squared loss and the logistic loss are convex and differentiable.

Safe screening rules make it possible to solve such problems more efficiently by discarding features whose coefficients are guaranteed to be zero at the optimum, prior to using a solver. They usual rely on the dual formulation of Equation (B.1):

$$\widehat{\boldsymbol{\theta}}^{(\lambda)} = \underset{\boldsymbol{\theta} \in \Delta_X}{\arg\max}\, D_\lambda(\boldsymbol{\theta}) := - \sum_{i=1}^{n} f_i^*(-\lambda \theta_i), \tag{B.2}$$

where $f_i^* : \mathbb{R} \to \mathbb{R}$ is the Fenchel-Legendre transform of f_i, defined by $f_i^*(u) = \sup_{z \in \mathbb{R}} \langle z, u \rangle - f_i(z)$ and $\Delta_X \subset \mathbb{R}^n$ is defined by $\Delta_X = \left\{ \boldsymbol{\theta} \in \mathbb{R}^n : \forall g = 1, \ldots, G, \Omega_g^D(X^{(g)\top} \boldsymbol{\theta}) \leq 1 \right\}$, where $\Omega_g^D : \mathbb{R}^{p_g} \to \mathbb{R}$ is the conjugate norm of Ω_g, defined by $\Omega_g^D(\boldsymbol{u}) = \max_{\boldsymbol{z} in \mathbb{R}^{p_g} : \Omega_g(\boldsymbol{z}) \leq 1} \langle \boldsymbol{z}, \boldsymbol{u} \rangle$, and $X^{(g)} \in \mathbb{R}^{n \times p_g}$ is the design matrix X restricted to the features/columns in group g.

In our setting,

- $\Omega_g^D(\boldsymbol{u}) = \left\| \boldsymbol{\beta}^{(g)} \right\|_2$ and $\Omega^D(\boldsymbol{u}) = \max_{g=1,\ldots,G} \frac{1}{w_g} \left\| \boldsymbol{u}^{(g)} \right\|_2$.
- If one uses the squared loss, that is to say, $f_i(z) = \frac{1}{2}(y_i - z)^2$, then $f_i^*(z) = \frac{1}{2} z^2 + y_i z$ and the Lipschitz constant is $\gamma = 1$.
- If one uses the logistic loss, that is to say, $\boldsymbol{y} \in \{0, 1\}^n$ and $f_i(z) = -y_i z + \log(1 + \exp(z))$, then
$$f_i^*(z) = \begin{cases} (z + y_i) \log(z + y_i) + (1 - (z + y_i)) \log(1 - (z + y_i)) & \text{if } 0 \leq (z + y_i) \leq 1 \\ +\infty & \text{otherwise,} \end{cases}$$
and the Lipschitz constant is $\gamma = 4$.

The general idea of safe screening rules, introduced by [EGVR10], is to find a region $\mathcal{R} \subset \mathbb{R}^n$ such that if $\widehat{\boldsymbol{\theta}}^{(\lambda)} \in \mathcal{R}$, for any $g \in \{1, \ldots, G\}$,

$$\Omega_g^D \left(X^{(g)\top} \widehat{\boldsymbol{\theta}}^{(\lambda)} \right) < 1 \Rightarrow \widehat{\boldsymbol{\beta}}^{(\lambda)} = 0.$$

Gap safe screening rules [N+17] exploit the duality gap $(P_\lambda(\boldsymbol{\beta}) - D_\lambda(\boldsymbol{\theta}))$ to obtain the radius of the safe region $\mathcal{R}$. More specifically, Ndiaye et al. show that $\forall \boldsymbol{\beta} \in \mathbb{R}^p, \forall \boldsymbol{\theta} \in \Delta_X$,

$$\left\| \widehat{\boldsymbol{\theta}}^{(\lambda)} - \boldsymbol{\theta} \right\|_2 \leq \sqrt{\frac{2(P_\lambda(\boldsymbol{\beta}) - D_\lambda(\boldsymbol{\theta}))}{\gamma \lambda^2}},$$

which leads them to define, for any $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\theta} \in \Delta_X$, the ball centered in $\boldsymbol{\theta}$ and of radius $\sqrt{\frac{2P_\lambda(\boldsymbol{\beta})-D_\lambda(\boldsymbol{\theta})}{\gamma\lambda^2}}$ as a safe region, that is to say a region that is guaranteed to contain $\widehat{\boldsymbol{\theta}}^{(\lambda)}$.

Appendix B.4. *Measuring selection stability*

To measure the stability of a feature selection property, we use the sample's Pearson coefficient [NB16]. This stability index is closely related to that proposed by Kuncheva [Kun08] and is appropriate for the comparison of feature sets of different sizes. This index relies on repeating the feature selection procedure M time (in this work, $M = 10$) and evaluating the overlap if the M resulting feature sets.

Each of the M sets of selected features can be represented by an indicator vector $\boldsymbol{s} \in \{0, 1\}^p$, where $s_j = 1$ if feature j is selected and 0 otherwise. The stability index between two feature sets $\mathcal{S}$ and $\mathcal{S}'$, represented by their indicator vectors $\boldsymbol{s}$ and $\boldsymbol{s}'$, is computed as the Pearsons's correlation between these two vectors:

$$\phi(\mathcal{S}, \mathcal{S}') = \frac{\sum_{j=1}^p (s_j - \bar{\boldsymbol{s}})(s'_j - \bar{\boldsymbol{s}'})}{\sqrt{\sum_{j=1}^p (s_j - \bar{\boldsymbol{s}})^2}\sqrt{\sum_{j=1}^p (s'_j - \bar{\boldsymbol{s}'})^2}}, \tag{B.3}$$

where $\bar{\boldsymbol{s}} = \frac{1}{p}\sum_{j=1}^p s_j$ and $\bar{\boldsymbol{s}'} = \frac{1}{p}\sum_{j=1}^p s'_j$.

Note that, because $\sum_{j=1}^p s_j = |\mathcal{S}|$, $\sum_{j=1}^p s_j s'_j = |\mathcal{S} \cap \mathcal{S}'|$, and $s_j^2 = s_j$, we can rewrite Equation (B.3) as

$$\phi(\mathcal{S}, \mathcal{S}') = \frac{|\mathcal{S} \cap \mathcal{S}'| - \frac{1}{p}|\mathcal{S}|\,|\mathcal{S}'|}{\sqrt{|\mathcal{S}|\left(1 - \frac{|\mathcal{S}|}{p}\right)}\sqrt{|\mathcal{S}'|\left(1 - \frac{|\mathcal{S}'|}{p}\right)}},$$

hence interpreting this index as the size of the intersection of the two sets, corrected by chance, that is to say, ensuring that the expected value of the index is 0 when the two selections are random.

The stability index between M sets of selected features is computed as the average pairwise stability index between all possible pairs of sets of selected features:

$$\phi(\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_M) = \frac{M(M-1)}{2}\sum_{k=1}^{M}\sum_{l=k+1}^{M}\phi(\mathcal{S}_k, \mathcal{S}_l). \tag{B.4}$$

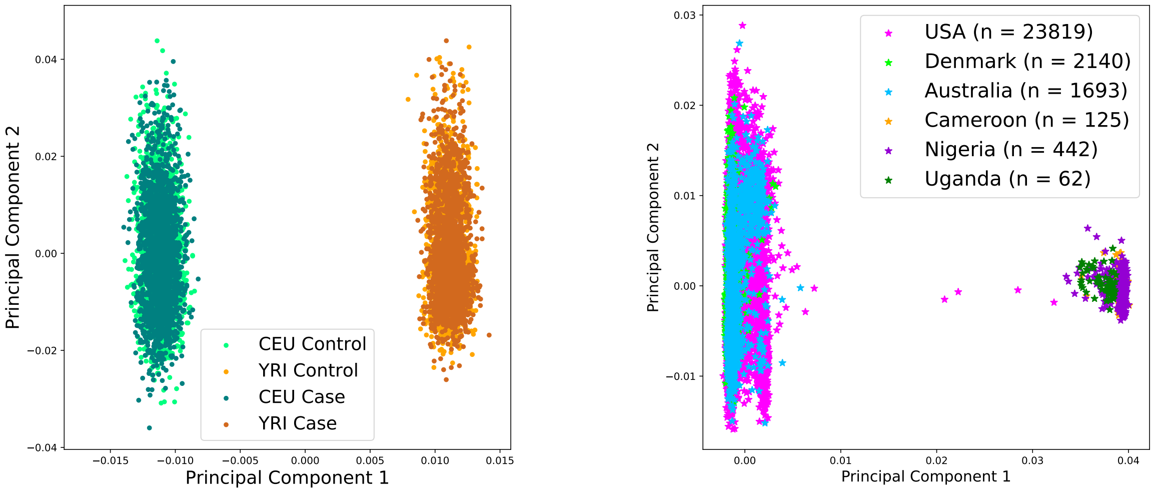
Appendix C. Supplementary Results

Appendix C.1. *PCA of the genotypes*

Figure C1 shows the genotypes of the simulated data (Figure C1a) and the DRIVE data (Figure C1b) projected on the two first principal components of the data.

Appendix C.2. *Runtimes*

Figure C2 shows the runtimes of the different Lasso methods on simulated data.



(a) Population structure in simulated data

(b) Population structure in the DRIVE data

Fig. C1. PCA for simulated and real datasets

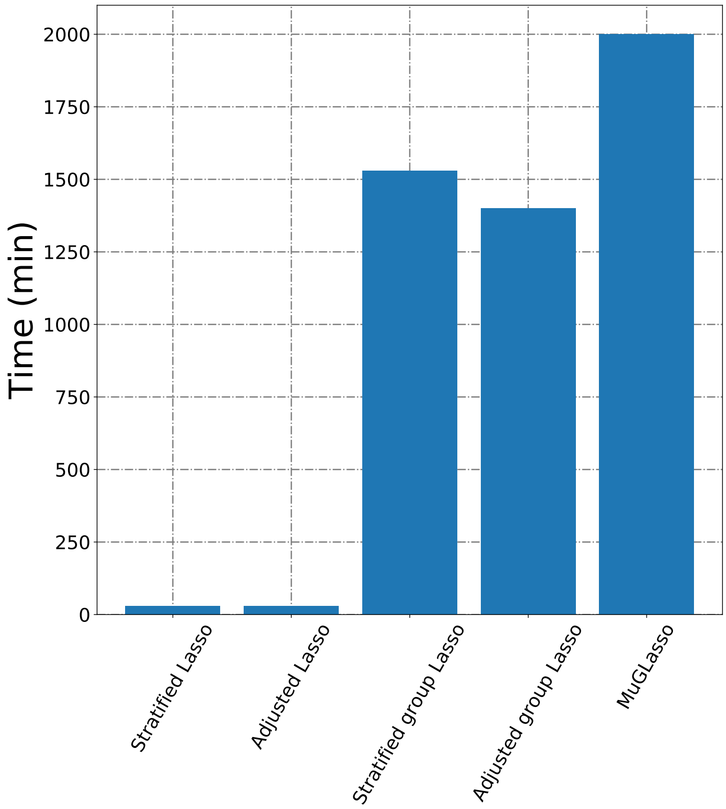


Fig. C2. Runtimes of the different Lasso approaches.

Appendix C.3. *Breast cancer risk loci detected by MuGLasso on DRIVE*

On the DRIVE dataset, MuGLasso selected 1 357 SNPs, forming 62 LD groups. Those SNPs include all the 306 SNPs that are significant in the adjusted GWAS approach. We used FUMA [WTVBP17] to analyze the remaining 1 051 SNPs, and found that 57% of these SNPs are within 10kb of protein coding genes. Hence MuGLasso identifies a total of 32 genes (listed in in Table C1), in addition to the 9 genes (*ITPR1*, *MRPS30*, *MAP3K1*, *SETD9*, *MIER3*, *EBF1*, *FGFR2*, *TOX3* and *MKL1*) identified by the adjusted GWAS.

Out of these 32 genes, 17 were previously identified in breast cancer meta-analyses which data include our 28 281 samples from the General Research Use dataset of the DRIVE Breast Cancer OncoArray Genotypes (see Table C1). More specifically, these studies respectively used 10 707 ER-negative breast cancer cases 76 649 controls [GC$^+$13] 45 290 cases and 41 880 controls of European ancestry [M$^+$13], 62 623 breast cancer cases and 61 696 controls [M$^+$15], 122 977 cases and 105 974 controls of European ancestry together with 14 068 cases and 13 104 controls of East Asian ancestry [M$^+$17a], and 210 088 controls (9 494 of which are BRCA1 mutation carriers) and 30 882 cases (21 468 ER-negative cases and 9 414 BRCA1 mutation carriers), all of European origin [M$^+$17].

This suggests that MuGLasso was able to rescue loci that are significant in a better-powered study (that is to say, a study with a larger number of samples).

In addition, we were able to find in the literature prior evidence of relationship with breast cancer risk or tumor growth for 7 additional genes, suggesting biological relevance of the MuGLasso findings.

Further analyses would be required to really get to the biological interpretation of these results. In particular, we restricted ourselves to mapping SNPs to genes based on a 10kb window, where other authors rather use 50kb, and FUMA provides many additional possibilities using known eQTLs and chromatin interactions across all tissues or for relevant tissues. In addition, pathway enrichment analyses could also be very relevant. One could also compare the selected SNPs to those significant in large meta-analyses such as [M$^+$17,M$^+$17a] in a more systematic manner to investigate how much power is gained by using MuGLasso on a subset of these GWAS data sets. Finally, we have analyzed jointly all selected SNPs and have not distinguished between those that are specific to one of the two populations and those that are common to both.

Table C1. The 32 potential breast cancer risk genes within 10kb of loci identified by MuGLasso and not the adjusted GWAS, together with information as to their biological relevance.

<table>
<tr><th colspan="2">Genes found in meta-GWAS including the samples used in this work</th></tr>
<tr><th>Gene symbols</th><th>Evidence</th></tr>
<tr><td><i>ASTN2</i></td><td>M^{+}17a</td></tr>
<tr><td><i>CCDC170</i></td><td>GC^{+}13,M^{+}13,M^{+}15,M^{+}17a,M^{+}17</td></tr>
<tr><td><i>CDYL2</i></td><td>M^{+}13,M^{+}15,M^{+}17a</td></tr>
<tr><td><i>DIRC3</i></td><td>M^{+}13,M^{+}15,M^{+}17a,M^{+}17</td></tr>
<tr><td><i>ELL</i></td><td>M^{+}13,M^{+}15,M^{+}17a,M^{+}17</td></tr>
<tr><td><i>ESR1</i></td><td>GC^{+}13,M^{+}15,M^{+}17a,M^{+}17</td></tr>
<tr><td><i>FTO</i></td><td>GC^{+}13,M^{+}13,M^{+}15,M^{+}17a,M^{+}17</td></tr>
<tr><td><i>GRHL1</i></td><td>M^{+}17a</td></tr>
<tr><td><i>KCNU1</i></td><td>M^{+}15,M^{+}17a</td></tr>
<tr><td><i>NEK10</i></td><td>M^{+}13,M^{+}15,M^{+}17a,M^{+}17</td></tr>
<tr><td><i>PAX9</i></td><td>M^{+}13,M^{+}15,M^{+}17a</td></tr>
<tr><td><i>PTHLH</i></td><td>GC^{+}13,M^{+}13,M^{+}15,M^{+}17a,M^{+}17</td></tr>
<tr><td><i>SSBP4</i></td><td>M^{+}17a</td></tr>
<tr><td><i>TGFBR2</i></td><td>M^{+}13,M^{+}15,M^{+}17a</td></tr>
<tr><td><i>TNRC6B</i></td><td>M^{+}17a</td></tr>
<tr><td><i>ZMIZ1</i></td><td>M^{+}13,M^{+}15,M^{+}17a</td></tr>
<tr><td><i>ZNF365</i></td><td>M^{+}17a,M^{+}17</td></tr>
<tr><th colspan="2">Genes found to be associated with breast cancer risk or tumor growth in the literature</th></tr>
<tr><th>Gene symbols</th><th>Evidence</th></tr>
<tr><td><i>ADSL</i></td><td>oncogenic driver in triple negative breast cancer [Z^{+}19]</td></tr>
<tr><td><i>CACNA1I</i></td><td>underexpressed in breast cancer [P^{+}17]</td></tr>
<tr><td><i>CCDC91</i></td><td>likely target gene of breast cancer risk variants [F^{+}19]</td></tr>
<tr><td><i>NUP205</i></td><td>forms a complex with NUP93 which regulates breast tumor growth [B^{+}20]</td></tr>
<tr><td><i>POP1</i></td><td>expression correlates with prognosis in breast cancer [L^{+}21]</td></tr>
<tr><td><i>PPFIBP1</i></td><td>promotes cell motility and migration in breast cancer [C^{+}16]</td></tr>
<tr><td><i>SGSM3</i></td><td>associated with breast cancer in a Chinese population [TZS17]</td></tr>
<tr><th colspan="2">Other genes</th></tr>
<tr><td colspan="2"><i>C7orf73</i>, <i>CCSER1</i>, <i>CD2AP</i>, <i>HK1</i>, <i>HRSP12</i>, <i>LUC7L3</i>, <i>MED21</i>, <i>REP15</i></td></tr>
</table>

Supplementary References

B^{+}20. Simone Bersini et al. Nup93 regulates breast tumor growth by modulating cell proliferation and actin cytoskeleton remodeling. *Life Sci Alliance*, 3(1), 2020.

C^{+}16. Sara Chiaretti et al. Effects of the scaffold proteins liprin-$\alpha 1$, $\beta 1$ and $\beta 2$ on invasion by breast cancer cells. *Biol Cell*, 108(3):65–75, 2016.

EGVR10. Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe feature elimination for the lasso and sparse supervised learning problems. *arXiv preprint arXiv:1009.4219*, 2010.

F^{+}19. Manuel A Ferreira et al. Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. *Nat Commun*, 10(1):1–18, 2019.

GC^{+}13. Montserrat Garcia-Closas et al. Genome-wide association studies identify four ER negative–specific breast cancer risk loci. *Nat Genet*, 45(4):392–398, 2013.

Kun08. Ludmila I. Kuncheva. A stability index for feature selection. *IASTED ICAIA*, 2008.

L^{+}21. Yang Liu et al. Identification of a three-RNA binding proteins (RBPs) signature predicting prognosis for breast cancer. *Front Oncol*, page 2150, 2021.

M$^+$13. Kyriaki Michailidou et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet*, 45(4):353–361, 2013.

M$^+$15. Kyriaki Michailidou et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat Genet*, 47(4):373–380, 2015.

M$^+$17a. Kyriaki Michailidou et al. Association analysis identifies 65 new breast cancer risk loci. *Nature*, 551(7678):92–94, 2017.

M$^+$17. Roger L Milne et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat Genet*, 49(12):1767–1778, 2017.

N$^+$17. Eugene Ndiaye et al. Gap safe screening rules for sparsity enforcing penalties. *Journal of Machine Learning Research 18*, 2017.

NB16. Sarah Nogueira and Gavin Brown. Measuring the stability of feature selection. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2016.

P$^+$17. Nam Nhut Phan et al. Voltage-gated calcium channels: Novel targets for cancer therapy. *Oncol Lett*, 14(2):2059–2074, 2017.

TZS17. Tan Tan, Kai Zhang, and Wenjun Chen Sun. Genetic variants of ESR1 and SGSM3 are associated with the susceptibility of breast cancer in the Chinese population. *Breast Cancer*, 24(3):369–374, 2017.

WTVBP17. Kyoko Watanabe, Erdogan Taskesen, Arjen Van Bochoven, and Danielle Posthuma. Functional mapping and annotation of genetic associations with fuma. *Nat Commun*, 8(1):1–11, 2017.

Z$^+$19. Giada Zurlo et al. Prolyl hydroxylase substrate adenylosuccinate lyase is an oncogenic driver in triple negative breast cancer. *Nat Commun*, 10(1):1–15, 2019.