



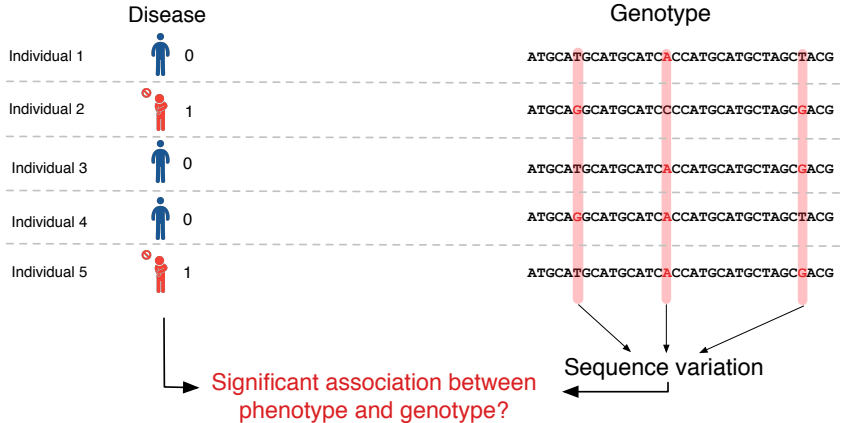
Network Mining in Biology and Medicine

Karsten Borgwardt

ETH Zürich, Department Biosystems

IBT Seminar, April 25, 2017

Mapping Phenotypes to the Genome



A **genome-wide association study (GWAS)** examines whether variation in the genome (in form of single nucleotide polymorphisms, SNPs) correlates with variation in the phenotype.

Missing Heritability

- Since 2001: More than 2000 new disease loci due to GWAS
- Problem: Phenotypic variance explained still disappointingly low

Vol 461|8 October 2009|doi:10.1038/nature08494

nature

REVIEWS

Finding the missing heritability of complex diseases

Teri A. Manolio¹, Francis S. Collins², Nancy J. Cox³, David B. Goldstein⁴, Lucia A. Hindorf⁵, David J. Hunter⁶, Mark I. McCarthy⁷, Erin M. Ramos⁵, Lon R. Cardon⁸, Aravinda Chakravarti⁹, Judy H. Cho¹⁰, Alan E. Guttmacher¹, Augustine Kong¹¹, Leonid Kruglyak¹², Elaine Mardis¹³, Charles N. Rotimi¹⁴, Montgomery Slatkin¹⁵, David Valle⁹, Alice S. Whittemore¹⁶, Michael Boehnke¹⁷, Andrew G. Clark¹⁸, Evan E. Eichler¹⁹, Greg Gibson²⁰, Jonathan L. Haines²¹, Trudy F. C. Mackay²², Steven A. McCarroll²³ & Peter M. Visscher²⁴

Genome-wide association studies have identified hundreds of genetic variants associated with complex human diseases and traits, and have provided valuable insights into their genetic architecture. Most variants identified so far confer relatively

Potential Reasons for Missing Heritability

Polygenic architectures

- Most current analyses neglect additive or multiplicative effects between loci → need for approaches for **multi-locus mapping**

Small effect sizes

- Not detectable with small sample sizes

Phenotypic effect of other genetic, epigenetic or non-genetic factors

- Genetic properties ignored so far, e.g. rare SNPs
- Epigenetic modifications of the genome
- Environmental effect on phenotype

Addressing Potential Reasons for Missing Heritability

Machine Learning in Genetics

- 1 Multi-locus mapping:
 - Algorithms to discover disease-related **systems of genetic loci**
- 2 Increasing sample size:
 - Algorithms that support **large-scale genotyping, association mapping and phenotyping**
- 3 Deciding whether additional information is required:
 - Tests that quantify the impact of **additional (epi)genetic factors**

Addressing Potential Reasons for Missing Heritability

Machine Learning in Genetics

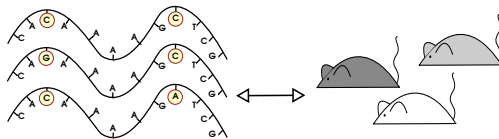
- 1 Multi-locus mapping:
 - Efficient algorithms for discovering disease-related SNP pairs (KDD 2011, ISMB 2011)
 - Efficient algorithms for discovering disease-related SNP sets (ISMB 2013, ISMB 2015, KDD 2015)
- 2 Increasing sample size:
 - Large-scale genotyping in *A. thaliana* (Nature Genetics 2011, Cell 2016)
 - Large-scale association mapping with mixed models (Bioinformatics 2013)
 - Automated image phenotyping of guppy fish (Bioinformatics 2011)
 - In silico phenotyping of migraine patients (Bioinformatics 2015)
- 3 Deciding whether additional information is required:
 - Assessing the stability of methylation across generations of *Arabidopsis* lab strains (Nature 2011a, PLoS Genetics 2015)
 - Assessing the functional impact of sequence variants (Human Mutation 2015)

Addressing Potential Reasons for Missing Heritability

Machine Learning in Genetics

- 1 Multi-locus mapping:
 - Efficient algorithms for discovering disease-related SNP pairs (KDD 2011, ISMB 2011)
 - Efficient algorithms for discovering disease-related SNP sets (ISMB 2013, ISMB 2015, KDD 2015)
- 2 Increasing sample size:
 - Large-scale genotyping in *A. thaliana* (Nature Genetics 2011, Cell 2016)
 - Large-scale association mapping with mixed models (Bioinformatics 2013)
 - Automated image phenotyping of guppy fish (Bioinformatics 2011)
- 3 Deciding whether additional information is required:
 - Assessing the stability of methylation across generations of *Arabidopsis* lab strains (Nature 2011a, PLoS Genetics 2015)
 - Assessing the functional impact of sequence variants (Human Mutation 2015)

Multi-Locus Models: Discovering Trait-Related Interactions



Problem statement

- Find the pair of SNPs most correlated with a binary phenotype

$$\operatorname{argmax}_{i,j} |r(\mathbf{x}_i, \mathbf{x}_j, \mathbf{y})|$$

- \mathbf{x}_i and \mathbf{x}_j represent one SNP each and \mathbf{y} is the phenotype; $\mathbf{x}_i, \mathbf{x}_j, \mathbf{y}$ are all n -dimensional vectors, given n individuals.
- There can be up to $p = 10^7$ SNPs, and order 10^{14} SNP pairs.
- Existing approaches: Greedy selection, Branch-and-bound strategies or index structures
→ low recall or worst-case $O(p^2)$ time

Multi-Locus Models: Discovering Trait-Related Interactions

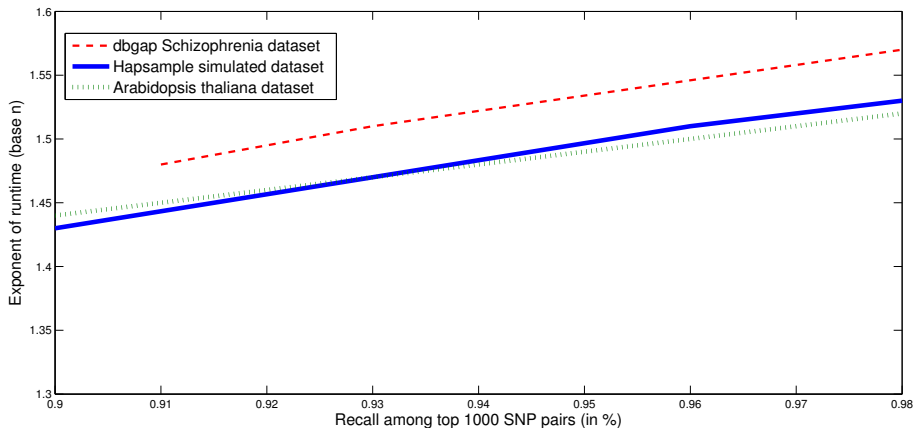
Lightbulb algorithm (Paturi et al., COLT 1989)

- Retrieves most correlated pair out of p binary vectors in $O(p^{1+\frac{\ln c_1}{\ln c_2}})$ via iterative hashing ($c_1 =$ highest, $c_2 =$ 2nd highest correlation)
- Discrepancy 1: SNPs are non-binary in general
- Discrepancy 2: Pearson's correlation coefficient

Our solution (Achlioptas et al., KDD 2011)

- Binarize our SNPs via locality sensitive hashing (Charikar, 2002).
- Show that the **Lightbulb algorithm on the binarized, transformed data** computes the solution to our **maximum correlation problem on the unbinarized original data**.
- Empirical result: Approx. $O(p^{1.5})$, speed-up of factor 1000 in practice

Multi-Locus Models: Discovering Trait-Related Interactions



Multi-Locus Models: Discovering Trait-Related Interactions

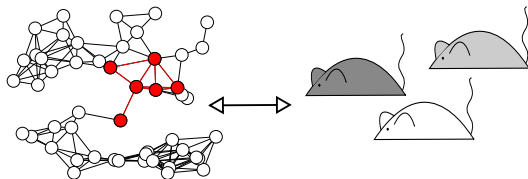
Alternative: Engineering approach

- Use parallel computing power of Graphical Processing Units for interaction discovery (Kam-Thong et al., ISMB 2011 & Human Heredity 2012)
- Similar speed-up as the Lightbulb algorithm

Road ahead

- We are part of the SNP \times SNP interaction discovery projects of
 - The international lung disease genetics consortium COPDGene
 - The international headache genetics consortium IHGC (Clinical Migraine)

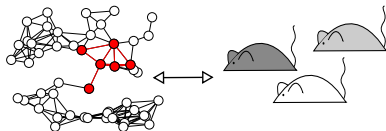
Multi-Locus Models: Discovering Trait-Related Networks



Network information

- What about models with more than 2 SNPs?
- Additive models are hard to interpret, multiplicative models are hard to compute.
- Can the growing knowledge about gene and protein networks be exploited to improve multi-locus mapping?

Multi-Locus Models: Discovering Trait-Related Networks



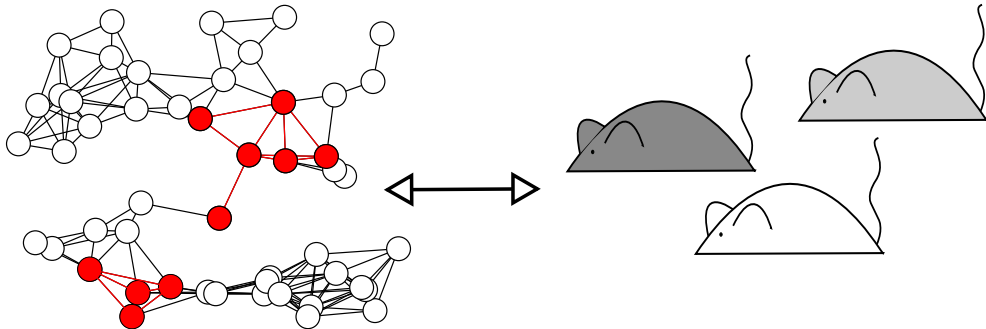
- Edges between SNPs near the same gene or SNPs in interacting genes
- c_i is the association score of SNP i , $f_i = 1$ if SNP i is selected, $f_i = 0$ if not.
- Find a set of SNPs with maximum total score:

$$\operatorname{argmax}_{\mathbf{f} \in \{0,1\}^p} \mathbf{c}^\top \mathbf{f}$$

such that

- the selected SNPs form a connected subgraph and
- \mathbf{f} is sparse.
- NP-complete problem: Maximum Weight Connected Subgraph Problem (Lee and Dooly, 1993)

Multi-Locus Models: Discovering Trait-Related Networks



Multi-Locus Models: Discovering Trait-Related Networks

Our formulation (Azencott et al., ISMB 2013)

- Networks are incomplete \rightarrow Connectedness needs not be strictly enforced, but merely rewarded by a Graph Laplacian regularizer $\mathbf{f}^\top \mathbf{L} \mathbf{f} = \sum_{i \sim j} (f_i - f_j)^2$, where $\mathbf{L} = \mathbf{D} - \mathbf{W}$.
- The SNP subnetwork selection problem is then:

$$\operatorname{argmax}_{\mathbf{f} \in \{0,1\}^p} \underbrace{\mathbf{c}^\top \mathbf{f}}_{\text{association}} - \underbrace{\lambda \mathbf{f}^\top \mathbf{L} \mathbf{f}}_{\text{connectivity}} - \underbrace{\eta \|\mathbf{f}\|_0}_{\text{sparsity}}$$

- This is a min-cut problem, for which efficient algorithms exist (we use Boykov and Kolmogorov, IEEE TPAMI 2004).
- Much faster and recovers four times more phenotype-related genes in *A. thaliana* than network-constrained Lasso models

Multi-Locus Models: Further Topics

Other important aspects

- Including prior knowledge on relevance of SNPs (Limin Li et al., ISMB 2011)
- Accounting for relatedness of individuals (Rakitsch et al., Bioinformatics 2013)
- Predicting multiple correlated phenotypes jointly (Rakitsch et al., NIPS 2013)
- **Measuring statistical significance**

Multi-Locus Mapping: Statistical Significance

Multiple Hypothesis Testing Problem

- What if we consider associations of groups of c SNPs with the phenotype?
- This leads to an enormous multiple testing problem: Any of the k SNP sets would correspond to a hypothesis that is tested ($k \in O(d^c)$).
- If unaccounted for, α per cent of all SNP sets might be considered significantly associated by random chance.
- It is imperative to control the family-wise error rate!
- If accounted for, e.g. by Bonferroni correction ($\frac{\alpha}{k}$), we might lose all statistical power.

Multi-Locus Mapping: Statistical Significance

Multiple Hypothesis Testing Problem

- What if we consider associations of groups of c SNPs with the phenotype?
- This leads to an **enormous multiple testing problem**: Any of the k SNP sets would correspond to a hypothesis that is tested ($k \in O(d^c)$).
- If unaccounted for, α per cent of all SNP sets might be considered significantly associated by random chance.
- It is imperative to control the **family-wise error rate!**
- If accounted for, e.g. by Bonferroni correction ($\frac{\alpha}{k}$), we might **lose all statistical power**.
- **Long considered unsolvable dilemma in Data Mining**
→ Starting Grant *Significant Pattern Mining (2015-2020)*

Multi-Locus Mapping: Statistical Significance

Tarone's trick

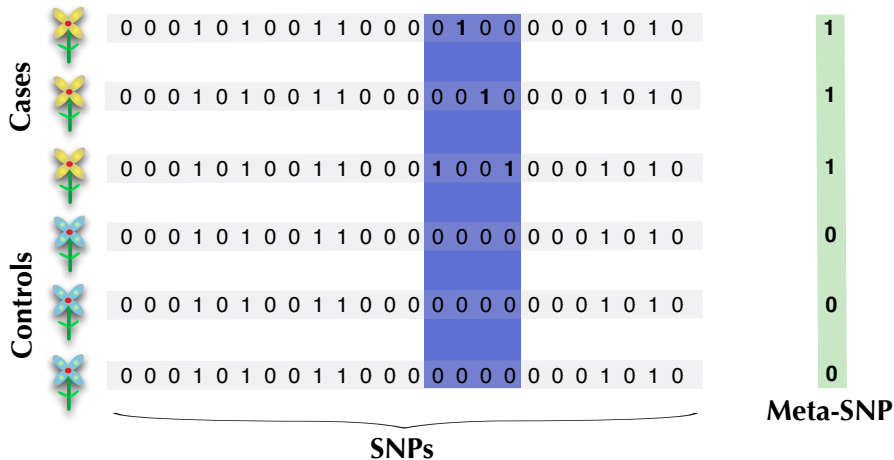
- Tarone (1990) noted that when working with discrete test statistics, e.g. Fisher's exact test, there is a **minimum p -value** that a feature combination can achieve.
- There are many **untestable hypotheses** whose minimum p -value is not smaller than $\frac{\alpha}{k}$.
- Only the remaining $m(k)$ **testable hypotheses** can reach significance at all.
- One can correct for $m(k)$ instead of k . As often $m(k) \ll k$, this greatly improves statistical power.
- **Grand data mining challenge: How to efficiently find $m(k)$ without running through all $k \in O(d^c)$ possible hypotheses?**
- We have developed frequent itemset mining algorithms for this task, which drastically improve the statistical power (SDM 2015, KDD 2015, ISMB 2015a, NIPS 2016, Bioinformatics 2017).

FAIS: Finding Intervals That Exhibit Genetic Heterogeneity

Genetic heterogeneity

- Genetic heterogeneity refers to the phenomenon that several different genes or sequence variants may give rise to the same phenotype.
- The correlation between each individual gene or variant and the phenotype may be too weak to be detected, but the group may have a strong correlation.
- The only current way to consider genetic heterogeneity is to consider fixed groups of variants. Genome-wide scans cause tremendous computational and statistical problems.

FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



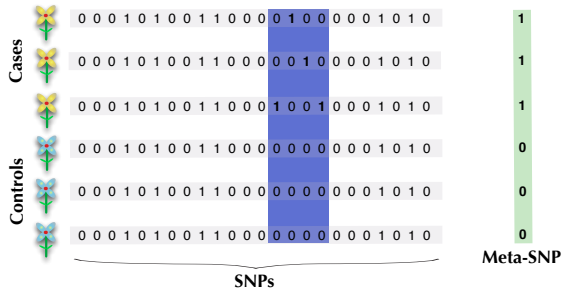
FAIS: Finding Intervals That Exhibit Genetic Heterogeneity

Fast Automatic Interval Search (Llinares-Lopez et al., ISMB 2015)

- Our goal is to **search for intervals that may exhibit genetic heterogeneity**, while
 - allowing for arbitrary start and end points of the intervals,
 - properly correcting for the inherent multiple testing problem, and
 - retaining statistical power and computational efficiency.
- We model the search as a **pattern mining problem**: Given an interval, an individual contains a pattern, if it has at least one minor allele in this interval.

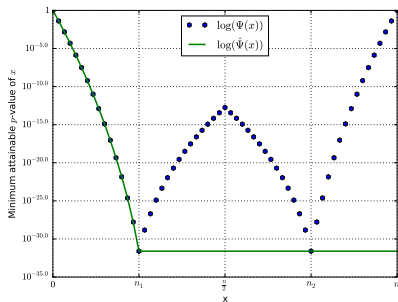
FAIS: Finding Intervals That Exhibit Genetic Heterogeneity

Finding trait-associated genome **segments** with at least one minor allele



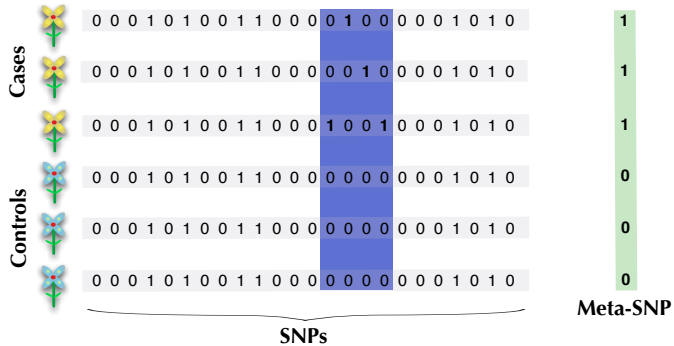
- An interval is represented by its maximum value. The longer an interval, the more likely it is that this maximum is 1.

FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



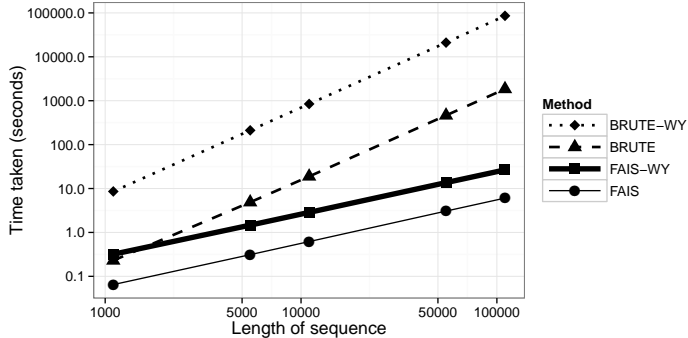
- **Pruning criterion 1:** If too many individuals have a particular pattern, the corresponding interval is not testable.

FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



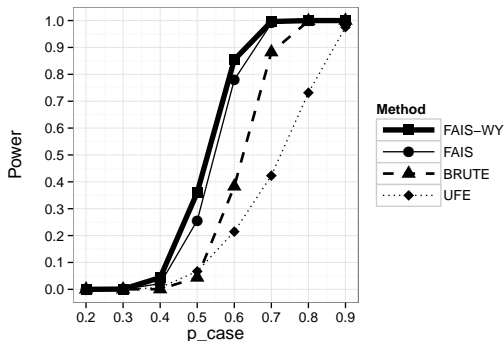
- **Pruning criterion 2:** If a pattern is too frequent to be testable, then none of the superintervals of the corresponding interval is testable.

FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



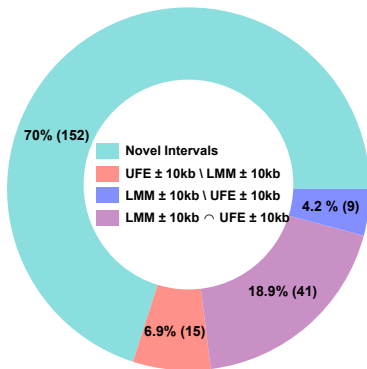
- Our method FAIS (Fast Automatic Interval Search) improves over the brute-force interval search in terms of runtime in simulations.

FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



- Our method FAIS (Fast Automatic Interval Search) improves over brute-force interval search and univariate approaches in terms of power in simulations.

FAIS: Finding Intervals That Exhibit Genetic Heterogeneity



- Most significant intervals would have been missed by univariate approaches (UFE and LMM) on 21 binary phenotypes from *Arabidopsis thaliana* (Atwell et al., Nature 2010).

FAIS: Conclusions and Outlook

Conclusions

- We can search for intervals that may exhibit genetic heterogeneity
 - efficiently,
 - without pre-defining the boundaries of intervals,
 - while properly correcting for multiple testing.

FAIS: Conclusions and Outlook

Conclusions

- We can search for intervals that may exhibit genetic heterogeneity
 - efficiently,
 - without pre-defining the boundaries of intervals,
 - while properly correcting for multiple testing.

Outlook: Genetic heterogeneity discovery

- How to account for covariates like age and gender? → Solution for categorial covariates (*NIPS 2016, Bioinformatics 2017*)
- How to extend our approach to non-binary encodings? → interpretation dependent
- How to extend our approach to networks of SNPs or genes? → current work

Some pointers

easyGWAS

- We have been developing easygwas.org, a Machine Learning platform for Geneticists (800 users in April 2017):



Software

Epistasis, SCONES, Significant Pattern Mining and Graph Kernels

- Epistasis: FAIS & FastCMH

www.bsse.ethz.ch/mlcb/research/bioinformatics-and-computational-biology/sis.html

www.bsse.ethz.ch/mlcb/research/bioinformatics-and-computational-biology/fastcmh.html

- Network GWAS: SCONES

www.bsse.ethz.ch/mlcb/research/bioinformatics-and-computational-biology/scones.html

- Significant Pattern Mining:

www.bsse.ethz.ch/mlcb/research/machine-learning/wylight.html

- Graph Kernels: www.bsse.ethz.ch/mlcb/research/machine-learning/graph-kernels

<http://www.bsse.ethz.ch/mlcb>

Network Mining in Biology and Medicine

Summary

- **Network Mining**, and more generally Multi-Locus Mapping, is a key technique to explore the genetic basis of complex traits.
- The high dimensionality of the problem leads to enormous **computational and statistical challenges**.
- **Solving both problems** at the same time is **largely unachieved**.
- We have developed several **Multi-Locus Mapping approaches** that **achieve both**.











Thank you










- Alfried-Krupp-Award for Young Professors
- Starting Grant (ERC-Backup Scheme of the SNSF)
- Horizon2020 Research and Innovation Action

<http://www.bsse.ethz.ch/mlcb>

References I

-  P. Achlioptas, *et al.* (ACM Press, 2011), p. 726.
-  C.-A. Azencott, *et al.*, *Bioinformatics* **29**, i171 (2013).
-  C. Becker, *et al.*, *Nature* **480**, 245 (2011).
-  J. Cao, *et al.*, *Nature Genetics* **43**, 956 (2011).
-  D. Grimm, *et al.*, *CoRR* [abs/1212.4788](#) (2012).
-  D. G. Grimm, *et al.*, *Human Mutation* **36**, 513 (2015).
-  J. Hagmann, *et al.*, *PLoS Genetics* **11**, e1004920 (2015).
-  T. Kam-Thong, *et al.*, *European Journal of Human Genetics* **19**, 465 (2011).
-  T. Kam-Thong, *et al.*, *Human Heredity* **73**, 220 (2012).
-  T. Karaletsos, *et al.*, *Bioinformatics* **28**, 1001 (2012).
-  L. Li, *et al.*, *Bioinformatics [ISMB/ECCB]* **27**, 342 (2011).

References II

-  F. Llinares-López, *et al.* (ACM Press, 2015), pp. 725–734.
-  F. Llinares-López, *et al.*, *Bioinformatics* **31**, i240 (2015).
-  L. Papaxanthos, *et al.*, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, *et al.*, eds. (2016), pp. 2271–2279.
-  B. Rakitsch, *et al.*, *Bioinformatics* **29**, 206 (2013).
-  B. Rakitsch, *et al.*, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, C. J. C. Burges, *et al.*, eds. (2013), pp. 1466–1474.
-  D. Roqueiro, *et al.*, *Bioinformatics* **31**, i303 (2015).
-  M. Sugiyama, *et al.*, *SIAM Data Mining*, S. Venkatasubramanian, J. Ye, eds. (SIAM, 2015), pp. 37–45.

Icon source: Icons made by Freepik from www.flaticon.com is licensed under CC BY 3.0