

## Evolutionäre Genomik

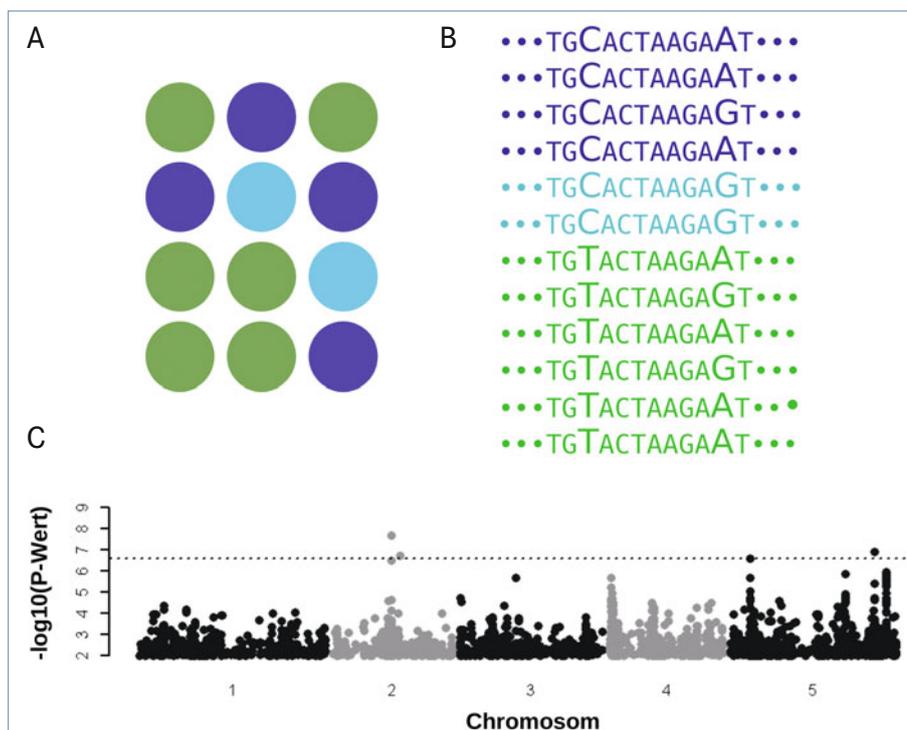
# Der Zusammenhang zwischen Genom und Phänotyp

ARTHUR KORTE

CENTER FOR COMPUTATIONAL AND THEORETICAL BIOLOGY (CCTB), UNIVERSITÄT WÜRZBURG

**Understanding the causal relationship between genotype and phenotype is a major objective in biology. Genome-wide association studies (GWAS) correlate genetic polymorphisms with trait variation and have already identified causative variants for various traits in many different organisms, from humans to plants. Importantly, many adaptive traits, like the regulation of flowering time in plants, are not regulated by distinct genetic effects, but by more sophisticated gene regulatory networks.**

DOI: 10.1007/s12268-022-1765-y  
© Der Autor 2022



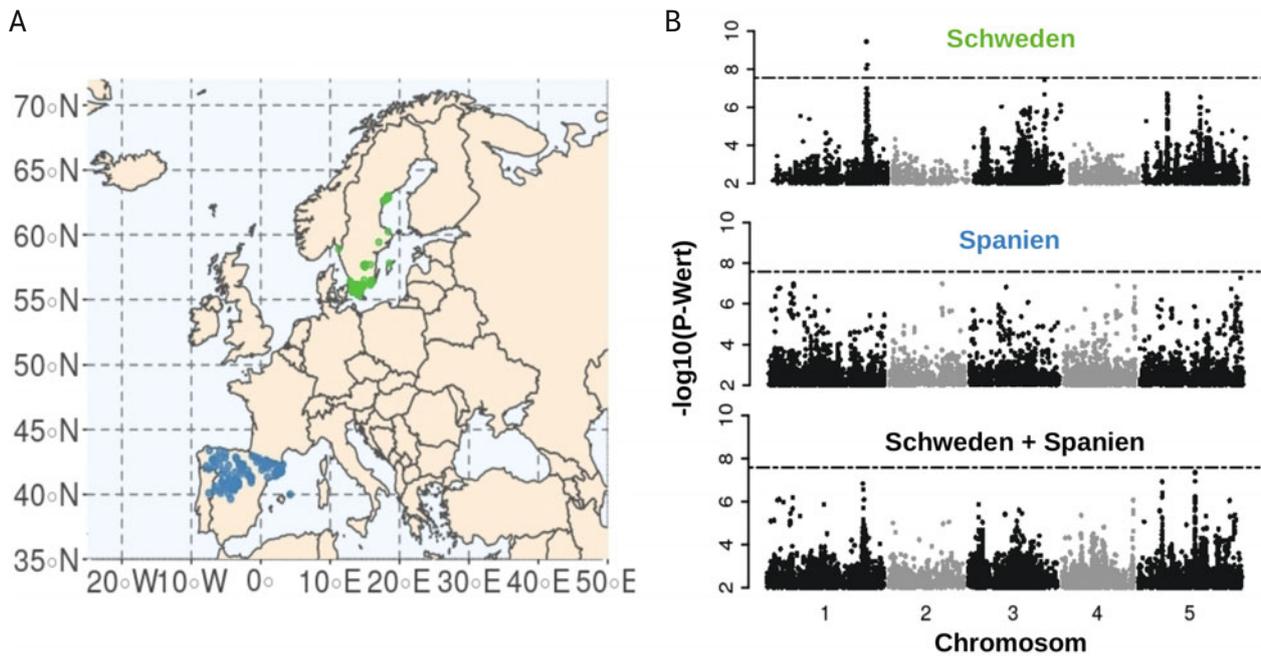
**▲ Abb. 1:** Hypothetisches Beispiel für eine GWAS mit dem Phänotyp „Grün“. **A**, 12 verschiedene Individuen sind als Punkte dargestellt und unterscheiden sich in ihrer Farbe. **B**, Ein Alignment eines Bereiches der sequenzierten Genome dieser Individuen. Die Genome der verschiedenen Individuen sind analog ihres Phänotyps farblich dargestellt. Die Genome unterscheiden sich im dargestellten Bereich an zwei Positionen (Position 3 und 11, fett). **C**, Manhattan-Plot der GWAS-Ergebnisse. Hierbei wird die Position jedes genomischen Markers gegen den negativen dekadischen Logarithmus des P-Wert des dazugehörigen Assoziationstests graphisch gegeneinander aufgetragen. Die horizontal gestrichelte Linie repräsentiert die Bonferroni-Schwelle. Assoziationen, die oberhalb dieser Linie sind, werden als statistisch signifikant bezeichnet.

■ Alle lebenden Organismen sehen unterschiedlich aus. Selbst Individuen der gleichen Art können sich in ihren Merkmalen unterscheiden. Diese phänotypischen Unterschiede können sowohl das äußere Erscheinungsbild als auch intrinsische, physiologische Merkmale betreffen. Unterschiede in Augen- oder Haarfarben gehen ebenso wie veränderte Blutzuckerwerte auf genomische Unterschiede des jeweiligen Individuums zurück. Man spricht hier von dem Zusammenhang zwischen Genotyp und Phänotyp. Dies bedeutet, dass Unterschiede im Genom eines Organismus zu Unterschieden im Erscheinungsbild führen können. Zusätzlich wird diese *genotype-phenotype map* durch externe Umwelteinflüsse moduliert. Das heißt, der Phänotyp eines Organismus ist durch sein Genom und die äußeren Umweltbedingungen determiniert.

Ein anschauliches Beispiel für den Einfluss der Umwelt ist die temperaturabhängige Veränderung von Farbe und Muster von Schmetterlingsflügeln. Hier zeigen genetische identische Individuen einen unterschiedlichen Phänotyp unter verschiedenen Temperaturen. Man spricht in diesem Zusammenhang auch von phänotypischer Plastizität, die einen großen Einfluss auf die Anpassungsfähigkeit von Organismen hat.

Die Beziehung von Genom und Phänotyp übt spätestens seit den Entdeckungen des Mönchs Gregor Mendel eine große Faszination auf die Wissenschaft aus. Mendel postulierte 1865 die Existenz „interner Faktoren“, die an die nächste Generation weitergegeben werden [1]. Auch wenn seine Arbeiten erst Jahrzehnte später – nach seinem Tod – allgemein gewürdigt wurden und Bateson den Begriff „Genetik“ erst am Anfang des 20. Jahrhunderts prägte [2], begründeten die Experimente von Mendel das Feld der klassischen Genetik und öffneten die Tür, um den Zusammenhang von Genotyp und Phänotyp zu verstehen.

Heutzutage ist dieser Zusammenhang von großem Interesse in verschiedenen Disziplinen, von der Evolutionsbiologie oder Molekularbiologie bis hin zur Medizin und der Land-



▲ **Abb. 2:** GWAS in verschiedenen Populationen der Modellpflanze *Arabidopsis thaliana*. **A,** Die geographische Herkunft der verschiedenen *Arabidopsis*-Individuen, die in der GWAS verwendet wurden, ist auf der Europakarte farblich gekennzeichnet. Es wurden Populationen aus Schweden und Spanien analysiert. **B,** Manhattan Plots der GWAS-Ergebnisse für die separate Analyse der schwedischen und spanischen Individuen, sowie die Ergebnisse der gemeinsamen Analyse aller Individuen. Der untersuchte Phänotyp war hier die Expression eines für den Blühzeitpunkt wichtigen Genes.

wirtschaft. Ziel ist es, für verschiedene Phänotypen die ursächlichen Gene oder genomischen Abschnitte zu identifizieren. Diese können dann experimentell validiert werden und schließlich z. B. als Marker für die Züchtung oder zur Entwicklung neuer Medikamente dienen.

### Wie findet man heraus, welche Gene für eine bestimmten Phänotyp wichtig sind ?

Das menschliche Genom besteht aus ungefähr drei Milliarden DNA-Basen. Auch wenn das Genom zwischen zwei Individuen zu 99,9 Prozent identisch ist, so unterscheiden sich zwei nicht näher verwandte Menschen doch in mehreren 100.000 Positionen im Genom. Viele dieser Unterschiede haben wahrscheinlich keinen Effekt, sie sind neutral. Einige dieser Unterschiede sind jedoch für unterschiedliche phänotypische Ausprägungen verantwortlich. Herauszufinden, welcher dieser Unterschiede jetzt mit einem bestimmten Phänotyp zusammenhängen ist nicht ganz trivial.

Um die ursächlichen kausalen genomischen Unterschiede zu identifizieren, wird die Methode der Genomweiten Assoziationsstudien (GWAS) benutzt. Diese Methode wurde 2005 zum ersten Mal publiziert und korreliert genomische Unterschiede innerhalb einer Population mit phänotypischen Unterschieden [3].

Zur Veranschaulichung ein abstraktes Beispiel: Stellen wir uns vor, dass wir an der genetischen Basis der Farbe eines Organismus interessiert sind. Dieser Organismus – in **Abbildung 1** als Punkt dargestellt – kommt natürlicherweise in drei verschiedenen, distinkten Farben vor: Grün, Lila und Hellblau. Wenn wir jetzt eruieren wollen, warum einige Individuen grün sind, bestimmen wir als erstes die Farbe in einer möglichst großen Population. Als nächstes brauchen wir die genomische Information aller Individuen. Durch moderne Sequenzieretechniken ist es heutzutage problemlos möglich, diese Information auch für tausende Individuen zu generieren. Nachdem wir nun das komplette Genom aller Individuen haben, müssen wir die verschiedenen Genome durch bioinformatische Analysen vergleichen und die genomischen Unterschiede identifizieren. In der Analyse konzentrieren wir uns auf *single-nucleotide polymorphisms* (SNPs), bei denen Individuen eine unterschiedliche DNA-Base an der selben Stelle im Genom aufweisen.

### Genomweite Assoziationsstudie (GWAS)

In einer GWAS wird jetzt für jeden dieser SNPs – die auch als genomischen Marker bezeichnet werden – ein statistischer Test durchgeführt. Vereinfacht dargestellt, wird getestet, ob die jeweiligen SNPs mit dem Phä-

notyp (in unserem Beispiel mit der Farbe Grün) korreliert.

In **Abbildung 1** sind exemplarisch zwölf Individuen und ein Teil ihres Genoms gezeigt. Das Genom der Individuen unterscheidet sich in dem gezeigten Bereich an zwei Stellen: An der Position 3 sind sowohl die Basen Cystein (C) und Thymin (T) in der Population vorhanden und an Position 11 treten die Basen Adenin (A) und Guanin (G) auf. Offensichtlich ist, dass alle grünen Individuen an Position 3 ein T haben und alle nicht grünen Individuen ein G. Hier gibt es eine Korrelation zwischen der Base an dieser Position und dem Phänotyp, wohingegen es für den Polymorphismus an Position 11 keinen Zusammenhang gibt.

Statistisch gesehen führen wir ein *Likelihood-ratio-test* für jede variable Position im Genom durch, um ein P-Wert für jeden Marker zu erzeugen. Diese P-Werte geben die Wahrscheinlichkeit an, ob ein etwaiger Marker mit dem beobachteten Phänotyp korreliert. Diese P-Werte werden dann schließlich in einen Manhattan-Plot graphisch dargestellt.

In marginaler Statistik wird meist eine Signifikanzwert (alpha) von 0,05 angenommen. Dieser Wert bedeutet, dass mit 95 Prozent Wahrscheinlichkeit die Werte nicht zufällig entstanden sind.

Jetzt ist es mit GWAS ein bisschen wie beim Lotto spielen: Je mehr Lose man kauft

(äquivalent zu : Je mehr Marker wir testen), desto höher ist die Wahrscheinlichkeit eines Gewinnes (einer signifikanten Assoziation). Deshalb müssen wir noch die Anzahl der Tests in unsere Berechnungen mit einbeziehen und penalisieren. Dies geschieht indem der marginale Signifikanzwert durch die Anzahl der zu testenden Marker geteilt wird. Wenn wir also 100.000 Marker testen, ist der Signifikanzwert =  $5 \cdot 10^{-7}$  und wir würden nur Werten, die noch signifikanter sind, vertrauen.

Das heißt, alle Marker, die unterhalb dieser Bonferroni-Schwelle liegen, sind statistisch signifikant mit dem Phänotyp assoziiert.

### Genetische Heterogenität

Mit der Methode des GWAS lassen sich also Gene, die für phänotypische Unterschiede wichtig sind, identifizieren. Es gibt jedoch ein paar zusätzliche Faktoren, die diese Analyse beeinflussen und die deshalb beachtet werden müssen.

Je mehr Individuen in einer Studie analysiert werden, desto höher sollte die Chance sein, signifikante Assoziationen zu identifizieren, da es mit mehr Individuen möglich ist, Marker mit einer geringeren Effektgröße statistisch nachzuweisen [4]. Diese Annahme hat aber nur dann Bestand, wenn die genetische Architektur in allen analysierten Individuen gleich ist. Dies ist jedoch nicht zwingend der Fall.

Wenn z. B. eine Population von Pflanzen einer Art an Standorten mit sehr unterschiedlichen klimatischen Bedingungen wachsen, kann die Anpassung an diese Bedingungen durch unterschiedliche Gene realisiert werden. Die Optimierung des Blühzeitpunktes in der Modellpflanze *Arabidopsis thaliana* wird z. B. durch unterschiedliche Gene in Populationen aus Spanien oder Schweden reguliert [5]. **Abbildung 2** zeigt die Ergebnisse eines GWAS für einen molekularen Phänotyp (der Expression eines Genes, diese Analyse wird auch als eGWAS bezeichnet) in geographisch verschiedenen Populationen. Das hier untersuchte Gen hat einen wichtigen Einfluss auf die Regulation des Blühzeitpunktes. Eine signifikante Assoziation kann hier nur in der schwedischen Population gefunden werden, nicht jedoch in der spanischen. Der Effekt des Markers unterscheidet sich in Ökotypen aus Spanien und Schweden. Wenn jetzt eine GWAS mit Pflanzen beider Populationen durchgeführt wird, kann es passieren, dass wichtige Asso-

ziationen zwischen Genotyp und Phänotyp nicht gefunden werden.

Diese genetische Heterogenität beschränkt sich nicht nur auf sehr unterschiedliche Populationen, wie die spanische und schwedische, sondern lässt sich auch beobachten, wenn näher verwandte Populationen analysiert werden. In einer GWAS von acht europäischen Populationen [5] sind von über 5.000 assoziierten SNPs mit einem marginalen P-Wert kleiner als  $10^{-4}$  nur acht SNPs in mehr als einer Population zu finden (**Abb. 3A**). Auch wenn wir die assoziierten Marker zu Regionen zusammenfassen, zeigt sich, dass die meisten Regionen nur in einer einzigen Population mit dem Blühzeitpunkt zusammenhängt (**Abb. 3B**). Dies bedeutet, dass die Regulation des Blühzeitpunktes in Populationen, die an unterschiedliche Umweltbedingungen angepasst sind, durch andere Gene und genomische Regionen geschieht.

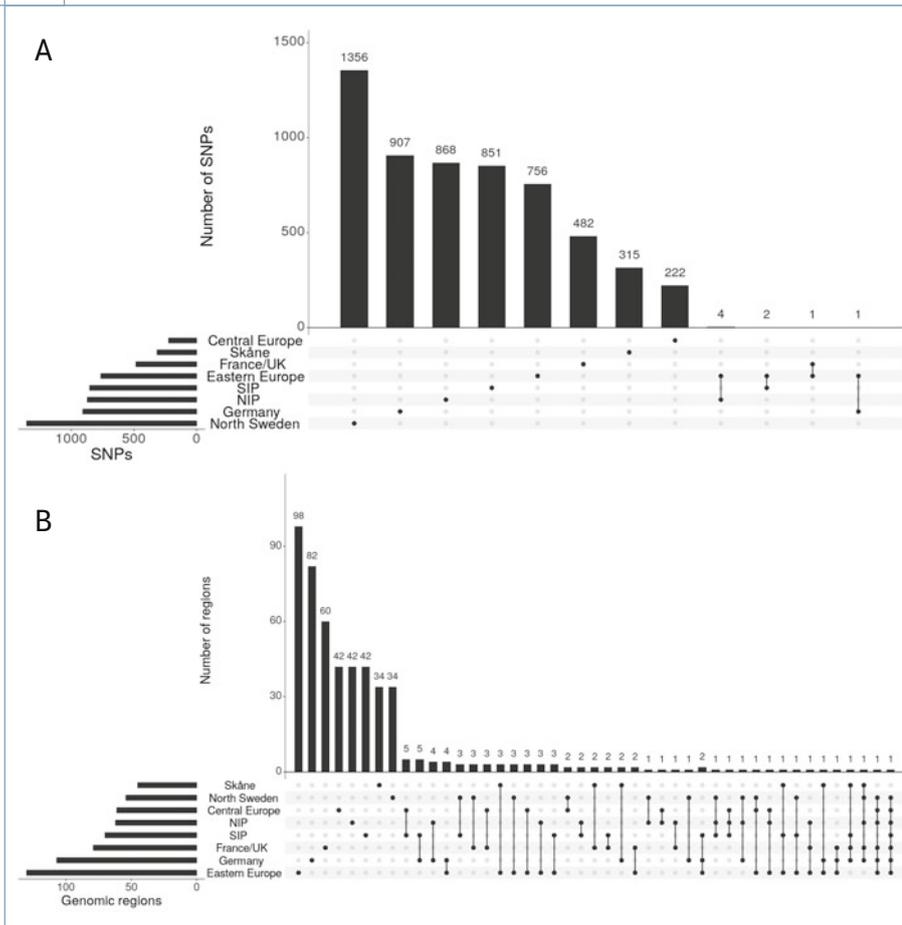
Dieses Problem der genetischen Heterogenität betrifft viele Phänotypen, insbesondere wenn diese unter einen Selektionsdruck stehen. So konnte in einer Studie gezeigt werden, dass die Zusammensetzung von bestimmten pflanzlichen Metaboliten (den Glucosinolaten), die für die Anpassung an biotischen und abiotischen Stress wichtig sind, sowohl auf lokaler als auch auf globaler Ebene eine große genetische Heterogenität aufweist [6]. Hier konnte gezeigt werden, dass eine Mischung von evolutionären Ereignissen für diese Heterogenität verantwortlich ist. Einerseits entstanden hier durch parallele Evolution Mutationen im selben Gen, die zu ähnlichen Merkmalsausprägungen führen. Andererseits konnten auch unabhängige Mutationen in verschiedenen Genen, die ebenfalls zu ähnlichen Merkmalen führen, beobachtet werden. In diesem Fall spricht man von konvergenter Evolution. Dieses Beispiel zeigt wie flexibel das Genom ist und sich an veränderte Umweltbedingungen anpassen kann: Ein Fakt, der es erschwert, signifikante Assoziationen zu finden.

### Gen-regulatorische Netzwerke

Um das ganze Szenario nochmals komplizierter zu machen, sind die meisten Phänotypen nicht das Produkt unabhängiger Effekte einzelner Gene, sondern der Kombination vieler Effekte und regulatorischer Netzwerke. Mendel hatte bei seinen Experimenten Phänotypen mit einer simplen Architektur gewählt: Die Form und Farbe

# Hier steht eine Anzeige.

 Springer



▲ **Abb. 3:** Ergebnisse einer GWAS für den Blühzeitpunkt in acht verschiedenen Populationen der Modellpflanze *Arabidopsis thaliana*. **A**, Histogramm der Anzahl an assoziierten SNPs (*P*-Wert kleiner als 10<sup>-4</sup>) in den acht verschiedenen Populationen und deren Überlappung. Die acht untersuchten Populationen sind durch ihre geographische Lage definiert. Dabei werden Populationen aus Zentraleuropa, Skåne (der südlichste Teil Schwedens), Frankreich und Großbritannien, Osteuropa, der südlichen spanischen Halbinsel (SIP), der nördlichen spanischen Halbinsel (NIP), Deutschland und Nordschweden unterschieden. **B**, Histogramm der Anzahl an assoziierten genomischen Regionen in den acht verschiedenen Populationen und deren Überlappung. Hier wurden SNPs die nicht weiter als 10 kb voneinander entfernt sind zu einer gemeinsamen genomischen Region zusammengefasst. Betrachtet wird die Überlappung von Regionen, auch wenn sich die zugrundeliegenden SNPs unterscheiden. Abbildung aus [5].

der Erbsen werden jeweils durch ein einzelnes Gen reguliert. Solche Phänotypen mit einfacher genetischer Architektur sind jedoch die Ausnahme. Es ist oft der Fall, dass der Effekt eines Gens abhängig von einem anderen ist. Diese epistatischen Interaktionen verkomplizieren das Finden von kausalen Genotyp-Phänotyp-Zusammenhängen, da diese nicht einfach in klassischen Modellen zu analysieren sind. Das Problem hier ist die schier unerschöpfliche Anzahl der Tests. Wenn wir alle paarweisen Interaktionen von einer Million Markern testen wollten, müssten wir eine Trillion statistische Tests durchführen. Dies ist sowohl statistisch als auch vom Rechenaufwand her ein Alptraum. Um diese Interaktionen zu identifizieren, ruht die Hoffnung auf Modellen des maschinellen Lernalers [7]. Die Idee ist, dass künstliche neuronale Netzwerke wichtige Interaktionen entdecken können ohne diese explizit zu modellieren.

### Vom GWAS zur Anwendung

Assoziationsstudien liefern wichtige Erkenntnisse und Hypothesen. Sie erlauben interessante Einblicke in die Evolution neuer Eigenschaften. GWAS erlaubt nicht nur die Identifikation kausativer Gene, sondern ermöglicht es, die Architektur von komplexen Phänotypen zu verstehen. Durch diese Einblicke können wir verstehen, wie die genetische Anpassung an lokale Gegebenheiten

ten und veränderte Umweltbedingungen funktioniert.

Die Anpassung von Pflanzen an veränderte Klimabedingungen basiert nicht auf dem Effekt einzelner Gene, sondern auf vielen subtilen Veränderungen in Gen-regulatorischen Netzwerken. Wichtig sind hierbei die *Hub*-Gene, die zentral ganze Netzwerke und deren Architektur modellieren. Das Identifizieren und Verstehen dieser spezifischen Netzwerke ist der Schlüssel, um Kulturpflanzen an den Klimawandel anzupassen. ■

### Literatur

- [1] Mendel G (1866) Versuche über Pflanzen-Hybriden. Verhandlungen des Naturforschenden Vereines in Brünn 4: 3-47
- [2] Bateson W (1909) Heredity and variation in modern lights. Darwin and Modern Science. Cambridge University Press
- [3] Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6: 95-108
- [4] Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9: 29
- [5] Lopez-Arboleda WA, Reinert S, Nordborg M et al. (2021) Global genetic heterogeneity in adaptive traits. Mol Biol Evol 38: 4822-483
- [6] Katz E, Li JJ, Jaegle B et al. (2021) Genetic variation, environment and demography intersect to shape *Arabidopsis* defense metabolite variation across Europe. eLife 10: e67784
- [7] Nicholls HL, John CR, Watson DS (2020) Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease Loci. Front Genet 11: 350

**Funding note:** Open Access funding enabled and organized by Projekt DEAL.  
**Open Access:** Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden. Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen. Weitere Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf <http://creativecommons.org/licenses/by/4.0/deed.de>.

### Korrespondenzadresse:

Dr. Arthur Korte  
 Evolutionary Genomics & Bioinformatics  
 Center for Computational and Theoretical Biology (CCTB)  
 Universität Würzburg  
 Am Hubland 32  
 D-97074 Würzburg  
 arthur.korte@uni-wuerzburg.de  
 www.cctb.uni-wuerzburg.de

### AUTOR



#### Arthur Korte

Jahrgang 1976. Biologiestudium an der Universität Freiburg. 2009 Promotion an der TU München. 2010–2015 Postdoc am Gregor-Mendel Institute of Molecular Plant Biology, Wien, Österreich. 2015–2021 Juniorprofessor für Evolutionäre Genomik am Center for Computational and Theoretical Biology, Universität Würzburg. Seit 2021 Arbeitsgruppenleiter Bioinformatics und Mitglied der kollegialen Leitung Center for Computational and Theoretical Biology, Universität Würzburg.