Artificial intelligence in human genomics

Image: © H. König, N. Matter-König | KIT

Evidence maps of research and of existing and emerging applications

The evidence maps were generated in the project DEEPEN GENOMICS ("Opportunities and challenges of the convergence of artificial intelligence, human genomics and genome editing"). The project (https://www.itas.kit.edu/english/projects_koen19_deepgen.php) ran from January 2019 until December 2020 and was funded within the framework "Innovations- und Technikanalyse" (Innovation and Technology Analysis), ITA of Germany's Federal Ministry of Education and Research, BMBF (https://www.bmbf.de/de/ita-vorhaben-10103.html).

The maps were developed to examine and present in an evidence-based manner (i.e., based on peer-reviewed scientific publications, conference papers/proceedings or patents) the current state of research as well as of associated existing and emerging applications.
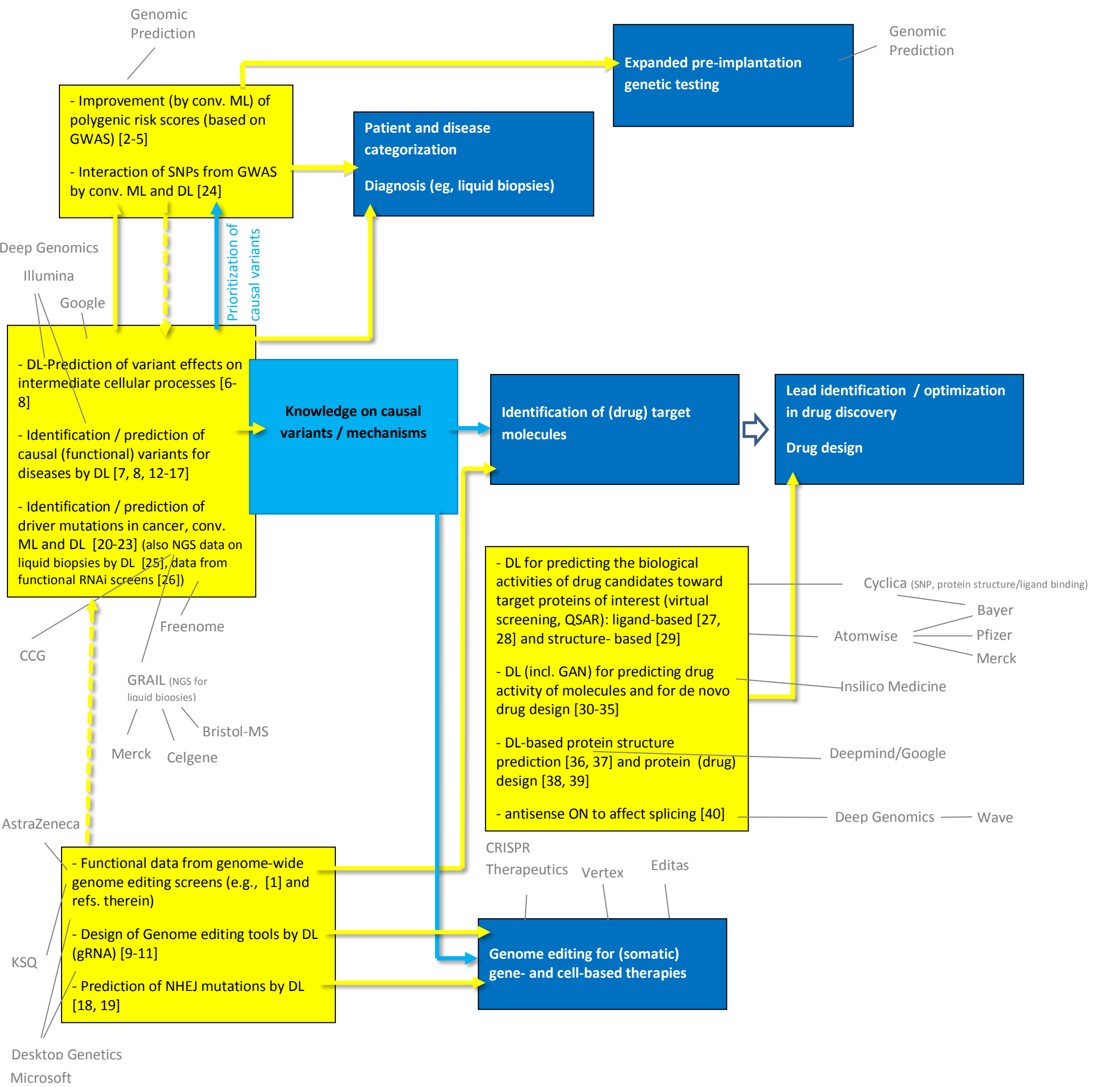

When using these maps, please refer to as:

König, Harald (2020): Artificial intelligence in human genomics: Mapping research and existing and emerging applications. Karlsruhe: Institute for Technology Assessment and Systems Analysis (KIT-ITAS)

Evidence map: Genomics-based (bio)medicine

Genomic Prediction

Genomic Prediction

Expanded pre-implantation genetic testing

- Improvement (by conv. ML) of polygenic risk scores (based on GWAS) [2-5]

- Interaction of SNPs from GWAS by conv. ML and DL [24]

Patient and disease categorization

Diagnosis (eg, liquid biopsies)

Prioritization of causal variants

Deep Genomics

Illumina

Google

- DL-Prediction of variant effects on intermediate cellular processes [6-8]

- Identification / prediction of causal (functional) variants for diseases by DL [7, 8, 12-17]

- Identification / prediction of driver mutations in cancer, conv. ML and DL [20-23] (also NGS data on liquid biopsies by DL [25], data from functional RNAi screens [26])

Knowledge on causal variants / mechanisms

Identification of (drug) target molecules

Lead identification / optimization in drug discovery

Drug design

CCG

Freenome

GRAIL (NGS for liquid biopsies)

Merck   Celgene   Bristol-MS

- DL for predicting the biological activities of drug candidates toward target proteins of interest (virtual screening, QSAR): ligand-based [27, 28] and structure- based [29]

- DL (incl. GAN) for predicting drug activity of molecules and for de novo drug design [30-35]

- DL-based protein structure prediction [36, 37] and protein (drug) design [38, 39]

- antisense ON to affect splicing [40]

Cyclica (SNP, protein structure/ligand binding)

Bayer

Atomwise   Pfizer

Merck

Insilico Medicine

Deepmind/Google

Deep Genomics   Wave

AstraZeneca

- Functional data from genome-wide genome editing screens (e.g., [1] and refs. therein)

- Design of Genome editing tools by DL (gRNA) [9-11]

- Prediction of NHEJ mutations by DL [18, 19]

CRISPR Therapeutics   Vertex   Editas

Genome editing for (somatic) gene- and cell-based therapies
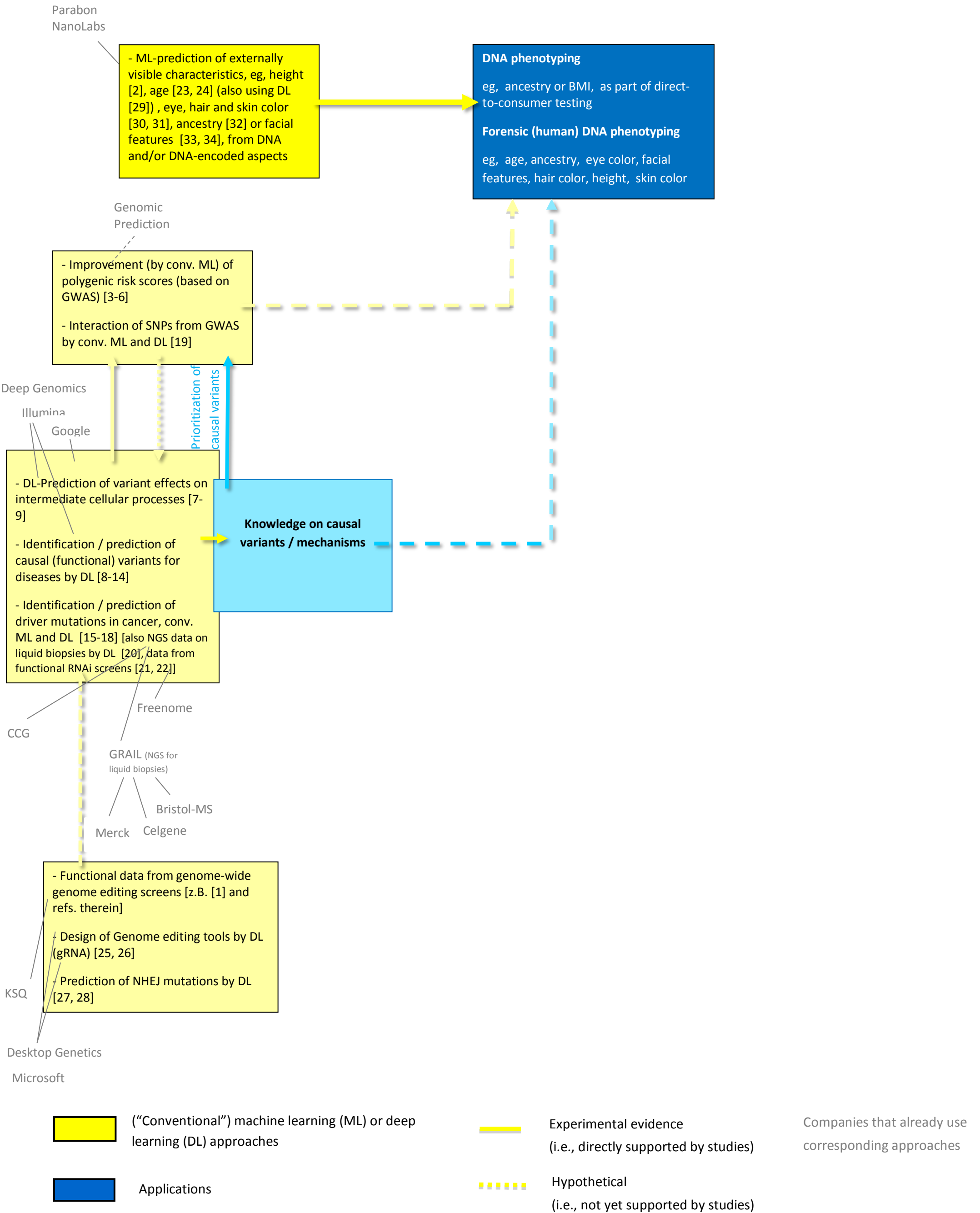
KSQ

Desktop Genetics
Microsoft

("Conventional") machine learning (ML) or deep learning (DL) approaches

Applications

Experimental evidence
(i.e., directly supported by studies)

Hypothetical
(i.e., not yet supported by studies)

Companies that already use corresponding approaches

References

1. Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., et al. (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell *163*, 1515-1526.
2. Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet *50*, 1219–1224
3. So, H.C., and Sham, P.C. (2017). Improving polygenic risk prediction from summary statistics by an empirical Bayes approach. Sci Rep-Uk *7*.
4. Paré, G., Mao, S., and Deng, W.Q. (2017). A machine-learning heuristic to improve gene score prediction of polygenic traits. Sci Rep-Uk *7*, 12665.
5. Lello, L., Raben, T.G., Yong, S.Y., Tellier, L., and Hsu, S.D.H. (2019). Genomic Prediction of 16 Complex Disease Risks Including Heart Attack, Diabetes, Breast and Prostate Cancer. Sci Rep *9*, 15286.
6. Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol *33*, 831-+.
7. Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y.M., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). The human splicing code reveals new insights into the genetic determinants of disease. Science *347*.
8. Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell *176*, 535-548 e524.
9. Listgarten, J., Weinstein, M., Kleinstiver, B.P., Sousa, A.A., Joung, J.K., Crawford, J., Gao, K., Hoang, L., Elibol, M., and Doench, J.G. (2018). Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. Nature Biomedical Engineering *2*, 38.
10. Kim, H.K., Min, S., Song, M., Jung, S., Choi, J.W., Kim, Y., Lee, S., Yoon, S., and Kim, H.H. (2018). Deep learning improves prediction of CRISPR–Cpf1 guide RNA activity. Nat Biotechnol *36*, 239.
11. Wang, D., Zhang, C., Wang, B., Li, B., Wang, Q., Liu, D., Wang, H., Zhou, Y., Shi, L., Lan, F., et al. (2019). Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. Nat Commun *10*, 4284.
12. Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods *12*, 931-934.
13. Zhou, J., Theesfeld, C.L., Yao, K., Chen, K.M., Wong, A.K., and Troyanskaya, O.G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat Genet *50*, 1171-1179.
14. Yao, V., Kaletsky, R., Keyes, W., Mor, D.E., Wong, A.K., Sohrabi, S., Murphy, C.T., and Troyanskaya, O.G. (2018). An integrative tissue-network approach to identify and test human disease genes. Nat Biotechnol *36*, 1091–1099.
15. Krishnan, A., Zhang, R., Yao, V., Theesfeld, C.L., Wong, A.K., Tadych, A., Volfovsky, N., Packer, A., Lash, A., and Troyanskaya, O.G. (2016). Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nat Neurosci *19*, 1454-1462.
16. Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F.C.P., Clarke, D., Gu, M., Emani, P., Yang, Y.T., et al. (2018). Comprehensive functional genomic resource and integrative model for the human brain. Science *362*.
17. Zhou, J., Park, C.Y., Theesfeld, C.L., Wong, A.K., Yuan, Y., Scheckel, C., Fak, J.J., Funk, J., Yao, K., Tajima, Y., et al. (2019). Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. Nat Genet *51*, 973-980.
18. Allen, F., Crepaldi, L., Alsinet, C., Strong, A.J., Kleshchevnikov, V., De Angeli, P., Páleníková, P., Khodak, A., Kiselev, V., Kosicki, M., et al. (2018). Predicting the mutations generated by repair of Cas9-induced double-strand breaks. Nat Biotechnol *37*, 64.
19. Shen, M.W., Arbab, M., Hsu, J.Y., Worstell, D., Culbertson, S.J., Krabbe, O., Cassa, C.A., Liu, D.R., Gifford, D.K., and Sherwood, R.I. (2018). Predictable and precise template-free CRISPR editing of pathogenic variants. Nature *563*, 646-651.
20. Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., et al. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell *174*, 1034-1035.
21. Luo, P., Ding, Y., Lei, X., and Wu, F.-X. (2019). deepDriver: Predicting Cancer Driver Genes Based on Somatic Mutations Using Deep Convolutional Neural Networks. Frontiers in genetics *10*.
22. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., and Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat Med *24*, 1559-1567.
23. Tokheim, C.J., Papadopoulos, N., Kinzler, K.W., Vogelstein, B., and Karchin, R. (2016). Evaluating the evaluation of cancer driver genes. Proc Natl Acad Sci U S A *113*, 14330-14335.
24. Uppu, S., Krishna, A., and Gopalan, R.P. (2018). A review on methods for detecting SNP interactions in high-dimensional genomic data. IEEE/ACM transactions on computational biology and bioinformatics *15*, 599-612.

25. Kothen-Hill, S.T., Zviran, A., Schulman, R.C., Deochand, S., Gaiti, F., Maloney, D., Huang, K.Y., Liao, W., Robine, N., and Omans, N.D. (2018). Deep learning mutation prediction enables early stage lung cancer detection in liquid biopsy.
26. Sanchez-Garcia, F., Villagrasa, P., Matsui, J., Kotliar, D., Castro, V., Akavia, U.D., Chen, B.J., Saucedo-Cuevas, L., Rodriguez Barrueco, R., Llobet-Navas, D., et al. (2014). Integration of genomic data enables selective discovery of breast cancer drivers. Cell *159*, 1461-1475.
27. Ma, J., Sheridan, R.P., Liaw, A., Dahl, G.E., and Svetnik, V. (2015). Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. Journal of Chemical Information and Modeling *55*, 263-274.
28. Unterthiner, T., Mayr, A., Klambauer, G., Steijaert, M., Wegner, J.K., Ceulemans, H., and Hochreiter, S. (2014). Deep learning as an opportunity in virtual screening. In Proceedings of the deep learning workshop at NIPS, Volume 27. pp. 1-9.
29. Wallach, I., Dzamba, M., and Heifets, A. (2015). AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXiv preprint arXiv:1510.02855.
30. Zhavoronkov, A., Ivanenkov, Y.A., Aliper, A., Veselov, M.S., Aladinskiy, V.A., Aladinskaya, A.V., Terentiev, V.A., Polykovskiy, D.A., Kuznetsov, M.D., Asadulaev, A., et al. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol *37*, 1038-1040.
31. Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A., and Zhavoronkov, A. (2017). druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. Molecular pharmaceutics *14*, 3098-3104.
32. Merk, D., Friedrich, L., Grisoni, F., and Schneider, G. (2018). De novo design of bioactive small molecules by artificial intelligence. Molecular informatics *37*, 1700153.
33. Merk, D., Grisoni, F., Friedrich, L., and Schneider, G. (2018). Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators. Communications Chemistry *1*, 68.
34. Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., Hernández-Lobato, J.M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T.D., Adams, R.P., and Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. ACS central science *4*, 268-276.
35. Stokes, J.M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N.M., MacNair, C.R., French, S., Carfrae, L.A., Bloom-Ackerman, Z., et al. (2020). A Deep Learning Approach to Antibiotic Discovery. Cell *180*, 688-702 e613.
36. Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Zidek, A., Nelson, A., Bridgland, A., and Penedones, H. (2018). De Novo structure prediction with deep-learning based scoring. Thirteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstracts).
37. Hou, J., Wu, T., Cao, R., and Cheng, J. (2019). Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. BioRxiv, 552422.
38. Silva, D.A., Yu, S., Ulge, U.Y., Spangler, J.B., Jude, K.M., Labao-Almeida, C., Ali, L.R., Quijano-Rubio, A., Ruterbusch, M., Leung, I., et al. (2019). De novo design of potent and selective mimics of IL-2 and IL-15. Nature *565*, 186-191.
39. Wang, J., Cao, H., Zhang, J.Z.H., and Qi, Y. (2018). Computational Protein Design with Deep Learning Neural Networks. Sci Rep-Uk *8*, 6349.
40. Merico, D., Spickett, C., O'Hara, M., Kakaradov, B., Deshwar, A., Fradkin, P., Gandhi, S., Gao, J., Grant, S., and Kron, K. (2019). ATP7B Variant c. 1934T> G p. Met645Arg Causes Wilson Disease by Promoting Exon 6 Skipping. bioRxiv, 693572.

Evidence map: DNA phenotyping / Forensic DNA phenotyping

Parabon
NanoLabs

- ML-prediction of externally visible characteristics, eg, height [2], age [23, 24] (also using DL [29]) , eye, hair and skin color [30, 31], ancestry [32] or facial features [33, 34], from DNA and/or DNA-encoded aspects

DNA phenotyping

eg, ancestry or BMI, as part of direct-to-consumer testing

Forensic (human) DNA phenotyping

eg, age, ancestry, eye color, facial features, hair color, height, skin color

Genomic
Prediction

- Improvement (by conv. ML) of polygenic risk scores (based on GWAS) [3-6]

- Interaction of SNPs from GWAS by conv. ML and DL [19]

Deep Genomics

Illumina

Google

Prioritization of causal variants

- DL-Prediction of variant effects on intermediate cellular processes [7-9]

- Identification / prediction of causal (functional) variants for diseases by DL [8-14]

- Identification / prediction of driver mutations in cancer, conv. ML and DL [15-18] [also NGS data on liquid biopsies by DL [20], data from functional RNAi screens [21, 22]]

Knowledge on causal variants / mechanisms

Freenome

CCG

GRAIL (NGS for liquid biopsies)

Bristol-MS

Merck   Celgene

- Functional data from genome-wide genome editing screens [z.B. [1] and refs. therein]

- Design of Genome editing tools by DL (gRNA) [25, 26]

- Prediction of NHEJ mutations by DL [27, 28]

KSQ

Desktop Genetics

Microsoft

("Conventional") machine learning (ML) or deep learning (DL) approaches

Applications

Experimental evidence
(i.e., directly supported by studies)

Hypothetical
(i.e., not yet supported by studies)

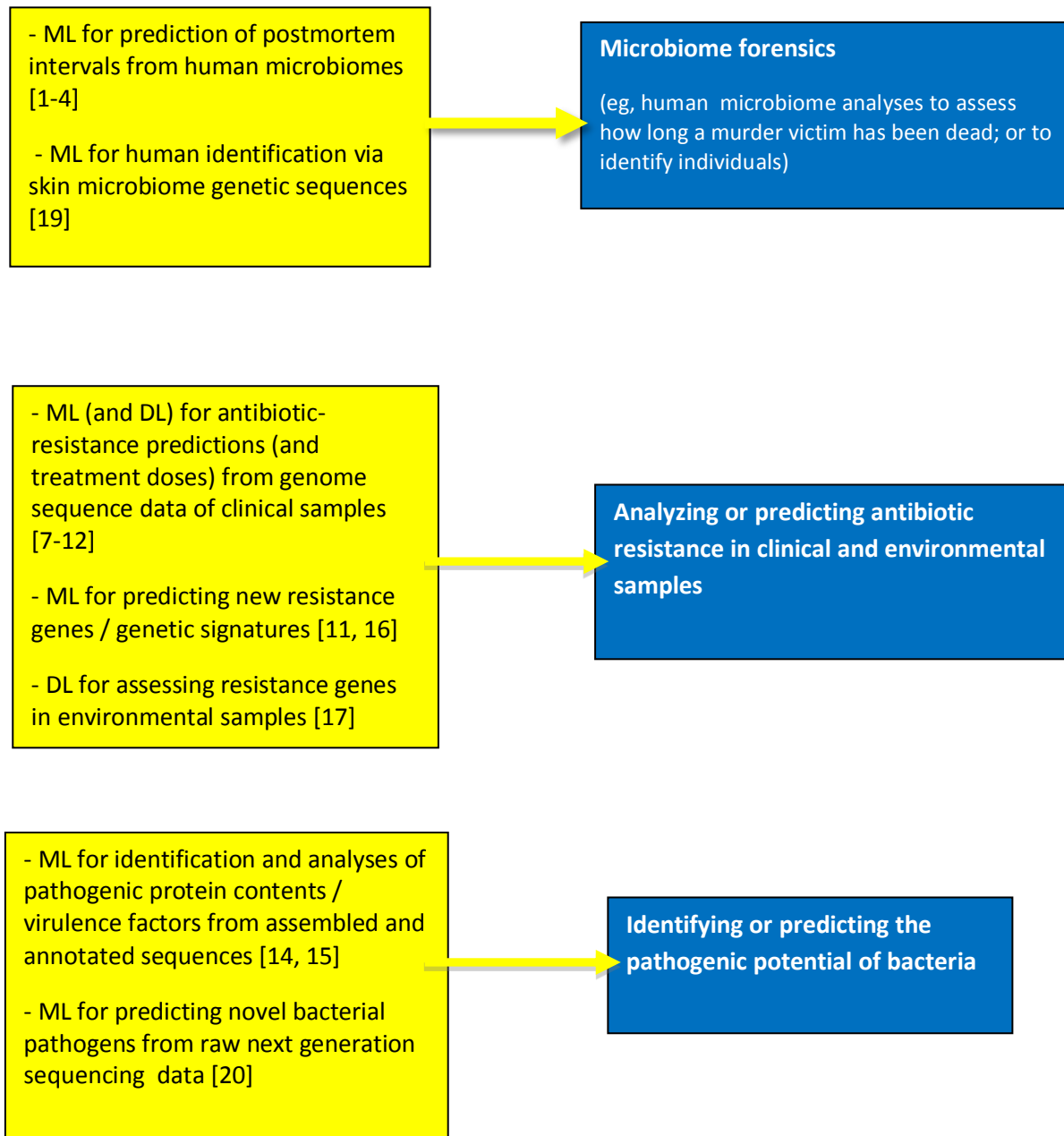Companies that already use corresponding approaches

References

1.   Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., et al. (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell *163*, 1515-1526.
2.   Lello, L., Avery, S.G., Tellier, L., Vazquez, A.I., de Los Campos, G., and Hsu, S.D.H. (2018). Accurate Genomic Prediction of Human Height. Genetics *210*, 477-497.
3.   Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet *50*, 1219–1224
4.   Lello, L., Raben, T.G., Yong, S.Y., Tellier, L.C., and Hsu, S.D. (2019). Genomic Prediction of Complex Disease Risk. BioRxiv, 506600.
5.   So, H.C., and Sham, P.C. (2017). Improving polygenic risk prediction from summary statistics by an empirical Bayes approach. Sci Rep-Uk *7*.
6.   Paré, G., Mao, S., and Deng, W.Q. (2017). A machine-learning heuristic to improve gene score prediction of polygenic traits. Sci Rep-Uk *7*, 12665.
7.   Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol *33*, 831-+.
8.   Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y.M., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). The human splicing code reveals new insights into the genetic determinants of disease. Science *347*.
9.   Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell *176*, 535-548 e524.
10.   Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods *12*, 931-934.
11.   Zhou, J., Theesfeld, C.L., Yao, K., Chen, K.M., Wong, A.K., and Troyanskaya, O.G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat Genet *50*, 1171-1179.
12.   Yao, V., Kaletsky, R., Keyes, W., Mor, D.E., Wong, A.K., Sohrabi, S., Murphy, C.T., and Troyanskaya, O.G. (2018). An integrative tissue-network approach to identify and test human disease genes. Nat Biotechnol *36*, 1091–1099.
13.   Krishnan, A., Zhang, R., Yao, V., Theesfeld, C.L., Wong, A.K., Tadych, A., Volfovsky, N., Packer, A., Lash, A., and Troyanskaya, O.G. (2016). Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nat Neurosci *19*, 1454-1462.
14.   Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F.C.P., Clarke, D., Gu, M., Emani, P., Yang, Y.T., et al. (2018). Comprehensive functional genomic resource and integrative model for the human brain. Science *362*.
15.   Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., et al. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell *174*, 1034-1035.
16.   Luo, P., Ding, Y., Lei, X., and Wu, F.-X. (2019). deepDriver: Predicting Cancer Driver Genes Based on Somatic Mutations Using Deep Convolutional Neural Networks. Frontiers in genetics *10*.
17.   Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., and Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat Med *24*, 1559-1567.
18.   Tokheim, C.J., Papadopoulos, N., Kinzler, K.W., Vogelstein, B., and Karchin, R. (2016). Evaluating the evaluation of cancer driver genes. Proc Natl Acad Sci U S A *113*, 14330-14335.
19.   Uppu, S., Krishna, A., and Gopalan, R.P. (2018). A review on methods for detecting SNP interactions in high-dimensional genomic data. IEEE/ACM transactions on computational biology and bioinformatics *15*, 599-612.
20.   Kothen-Hill, S.T., Zviran, A., Schulman, R.C., Deochand, S., Gaiti, F., Maloney, D., Huang, K.Y., Liao, W., Robine, N., and Omans, N.D. (2018). Deep learning mutation prediction enables early stage lung cancer detection in liquid biopsy.
21.   Sanchez-Garcia, F., Villagrasa, P., Matsui, J., Kotliar, D., Castro, V., Akavia, U.D., Chen, B.J., Saucedo-Cuevas, L., Rodriguez Barrueco, R., Llobet-Navas, D., et al. (2014). Integration of genomic data enables selective discovery of breast cancer drivers. Cell *159*, 1461-1475.
22.   Pechal, J.L., Crippen, T.L., Tarone, A.M., Lewis, A.J., Tomberlin, J.K., and Benbow, M.E. (2013). Microbial community functional change during vertebrate carrion decomposition. PloS one *8*, e79035.
23.   Hong, S.R., Jung, S.-E., Lee, E.H., Shin, K.-J., Yang, W.I., and Lee, H.Y. (2017). DNA methylation-based age prediction from saliva: high age predictability by combination of 7 CpG markers. Forensic Science International: Genetics *29*, 118-125.
24.   Aliferi, A., Ballard, D., Gallidabino, M.D., Thurtle, H., Barron, L., and Court, D.S. (2018). DNA methylation-based age prediction using massively parallel sequencing data and multiple machine learning models. Forensic Science International: Genetics *37*, 215-226.
25.   Listgarten, J., Weinstein, M., Kleinstiver, B.P., Sousa, A.A., Joung, J.K., Crawford, J., Gao, K., Hoang, L., Elibol, M., and Doench, J.G. (2018). Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. Nature Biomedical Engineering *2*, 38.
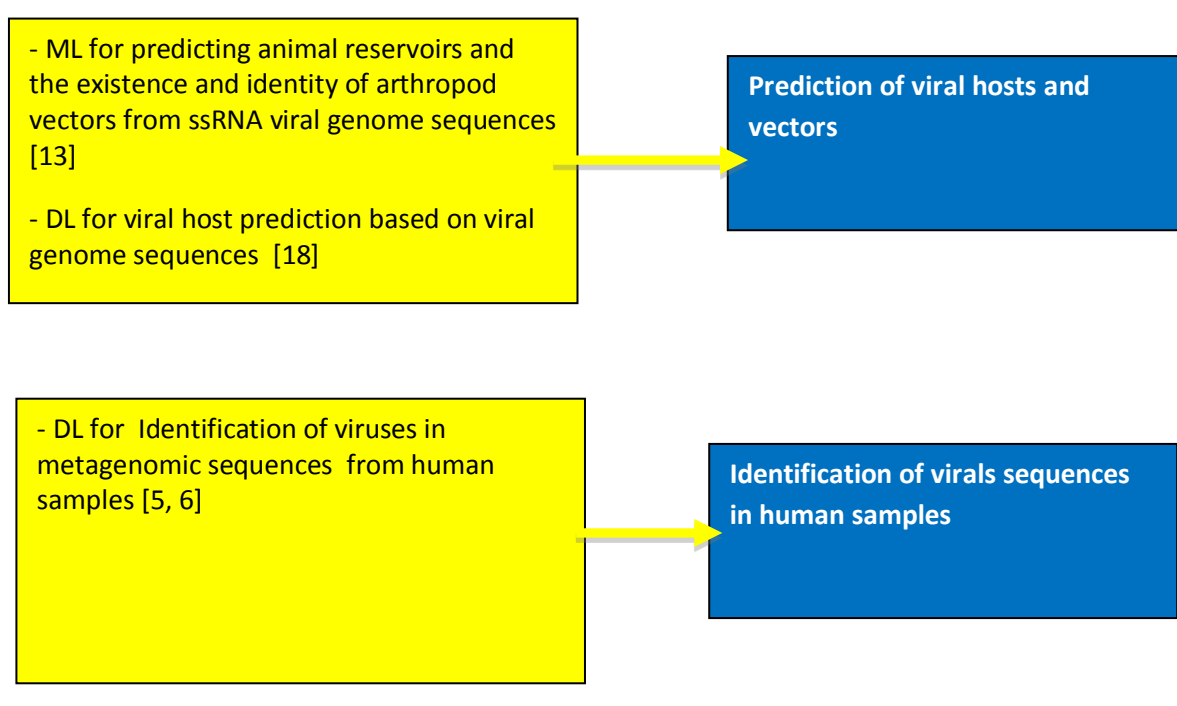
26. Kim, H.K., Min, S., Song, M., Jung, S., Choi, J.W., Kim, Y., Lee, S., Yoon, S., and Kim, H.H. (2018). Deep learning improves prediction of CRISPR–Cpf1 guide RNA activity. Nat Biotechnol *36*, 239.
27. Allen, F., Crepaldi, L., Alsinet, C., Strong, A.J., Kleshchevnikov, V., De Angeli, P., Páleníková, P., Khodak, A., Kiselev, V., Kosicki, M., et al. (2018). Predicting the mutations generated by repair of Cas9-induced double-strand breaks. Nat Biotechnol *37*, 64.
28. Shen, M.W., Arbab, M., Hsu, J.Y., Worstell, D., Culbertson, S.J., Krabbe, O., Cassa, C.A., Liu, D.R., Gifford, D.K., and Sherwood, R.I. (2018). Predictable and precise template-free CRISPR editing of pathogenic variants. Nature *563*, 646-651.
29. Vidaki, A., Ballard, D., Aliferi, A., Miller, T.H., Barron, L.P., and Court, D.S. (2017). DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. Forensic Science International: Genetics *28*, 225-236.
30. Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., Branicki, W., and Kayser, M. (2013). The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. Forensic Science International: Genetics *7*, 98-115.
31. Chaitanya, L., Breslin, K., Zuñiga, S., Wirken, L., Pośpiech, E., Kukla-Bartoszek, M., Sijen, T., de Knijff, P., Liu, F., and Branicki, W. (2018). The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation. Forensic Science International: Genetics *35*, 123-135.
32. Pereira, L., Alshamali, F., Andreassen, R., Ballard, R., Chantratita, W., Cho, N.S., Coudray, C., Dugoujon, J.-M., Espinoza, M., and González-Andrade, F. (2011). PopAffiliator: online calculator for individual affiliation to a major population group based on 17 autosomal short tandem repeat genotype profile. International journal of legal medicine *125*, 629-636.
33. Lippert, C., Sabatini, R., Maher, M.C., Kang, E.Y., Lee, S., Arikan, O., Harley, A., Bernal, A., Garst, P., Lavrenko, V., et al. (2017). Identification of individuals by trait prediction using whole-genome sequencing data. Proceedings of the National Academy of Sciences *114*, 10166-10171.
34. Sero, D., Zaidi, A., Li, J., White, J.D., Zarzar, T.B.G., Marazita, M.L., Weinberg, S.M., Suetens, P., Vandermeulen, D., Wagner, J.K., et al. (2019). Facial recognition from DNA using face-to-DNA classifiers. Nature Communications *10*, 2557.

Evidence map: Human metagenomics

Bacteria

- ML for prediction of postmortem intervals from human microbiomes [1-4]

- ML for human identification via skin microbiome genetic sequences [19]

Microbiome forensics

(eg, human microbiome analyses to assess how long a murder victim has been dead; or to identify individuals)

- ML (and DL) for antibiotic-resistance predictions (and treatment doses) from genome sequence data of clinical samples [7-12]

- ML for predicting new resistance genes / genetic signatures [11, 16]

- DL for assessing resistance genes in environmental samples [17]

Analyzing or predicting antibiotic resistance in clinical and environmental samples

---

- ML for identification and analyses of pathogenic protein contents / virulence factors from assembled and annotated sequences [14, 15]

- ML for predicting novel bacterial pathogens from raw next generation sequencing data [20]

Identifying or predicting the pathogenic potential of bacteria

Viruses

- ML for predicting animal reservoirs and the existence and identity of arthropod vectors from ssRNA viral genome sequences [13]

- DL for viral host prediction based on viral genome sequences [18]

Prediction of viral hosts and vectors

- DL for Identification of viruses in metagenomic sequences from human samples [5, 6]

Identification of virals sequences in human samples

("Conventional") machine learning (ML) or deep learning (DL) approaches

Applications

Experimental evidence
(i.e., directly supported by studies)

References

1. Johnson, H.R., Trinidad, D.D., Guzman, S., Khan, Z., Parziale, J.V., DeBruyn, J.M., and Lents, N.H. (2016). A machine learning approach for using the postmortem skin microbiome to estimate the postmortem interval. PloS one *11*, e0167370.
2. Pechal, J.L., Crippen, T.L., Tarone, A.M., Lewis, A.J., Tomberlin, J.K., and Benbow, M.E. (2013). Microbial community functional change during vertebrate carrion decomposition. PloS one *8*, e79035.
3. Belk, A., Xu, Z.Z., Carter, D.O., Lynne, A., Bucheli, S., Knight, R., and Metcalf, J. (2018). Microbiome data accurately predicts the postmortem interval using random forest regression models. Genes *9*, 104.
4. Zhang, Y., Pechal, J.L., Schmidt, C.J., Jordan, H.R., Wang, W.W., Benbow, M.E., Sze, S.-H., and Tarone, A.M. (2019). Machine learning performance in a microbial molecular autopsy context: A cross-sectional postmortem human population study. PloS one *14*, e0213829.
5. Tampuu, A., Bzhalava, Z., Dillner, J., and Vicente, R. (2019). ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples. PloS one *14*, e0222271.
6. Ren, J., Song, K., Deng, C., Ahlgren, N.A., Fuhrman, J.A., Li, Y., Xie, X., and Sun, F. (2018). Identifying viruses from metagenomic data by deep learning. arXiv preprint arXiv:1806.07810.
7. Bradley, P., Gordon, N.C., Walker, T.M., Dunn, L., Heys, S., Huang, B., Earle, S., Pankhurst, L.J., Anson, L., de Cesare, M., et al. (2015). Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. Nature Communications *6*, 10063.
8. Davis, J.J., Boisvert, S., Brettin, T., Kenyon, R.W., Mao, C., Olson, R., Overbeek, R., Santerre, J., Shukla, M., Wattam, A.R., et al. (2016). Antimicrobial Resistance Prediction in PATRIC and RAST. Sci Rep-Uk *6*, 27930.
9. Pesesky, M.W., Hussain, T., Wallace, M., Patel, S., Andleeb, S., Burnham, C.D., and Dantas, G. (2016). Evaluation of Machine Learning and Rules-Based Approaches for Predicting Antimicrobial Resistance Profiles in Gram-negative Bacilli from Whole Genome Sequence Data. Front Microbiol *7*, 1887.
10. Nguyen, M., Long, S.W., McDermott, P.F., Olsen, R.J., Olson, R., Stevens, R.L., Tyson, G.H., Zhao, S., and Davis, J.J. (2019). Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal Salmonella. Journal of clinical microbiology *57*, e01260-01218.
11. Kouchaki, S., Yang, Y., Walker, T.M., Sarah Walker, A., Wilson, D.J., Peto, T.E., Crook, D.W., Consortium, C., and Clifton, D.A. (2018). Application of machine learning techniques to tuberculosis drug resistance analysis. Bioinformatics *35*, 2276-2282.
12. Moradigaravand, D., Palm, M., Farewell, A., Mustonen, V., Warringer, J., and Parts, L. (2018). Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data. PLoS computational biology *14*, e1006258.
13. Babayan, S.A., Orton, R.J., and Streicker, D.G. (2018). Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. Science *362*, 577-580.
14. Iraola, G., Vazquez, G., Spangenberg, L., and Naya, H. (2012). Reduced set of virulence genes allows high accuracy prediction of bacterial pathogenicity in humans. PloS one *7*, e42144.
15. Cosentino, S., Voldby Larsen, M., Moller Aarestrup, F., and Lund, O. (2013). PathogenFinder--distinguishing friend from foe using bacterial whole genome sequence data. PloS one *8*, e77302.
16. Kavvas, E.S., Catoiu, E., Mih, N., Yurkovich, J.T., Seif, Y., Dillon, N., Heckmann, D., Anand, A., Yang, L., and Nizet, V. (2018). Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. Nature Communications *9*, 4306.
17. Arango-Argoty, G., Garner, E., Pruden, A., Heath, L.S., Vikesland, P., and Zhang, L. (2018). DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome *6*, 23.
18. Mock, F., Viehweger, A., Barth, E., and Marz, M. (2019). Viral host prediction with Deep Learning. BioRxiv, 575571.
19. Schmedes, S.E., Woerner, A.E., Novroski, N.M., Wendt, F.R., King, J.L., Stephens, K.M., and Budowle, B. (2018). Targeted sequencing of clade-specific markers from skin microbiomes for forensic human identification. Forensic Science International: Genetics *32*, 50-61.
20. Deneke, C., Rentzsch, R., and Renard, B.Y. (2017). PaPrBaG: A machine learning approach for the detection of novel pathogens from NGS data. Sci Rep-Uk *7*, 39194.