Darwin and Fisher meet at biotech: On the potential of computational molecular evolution in industry

Maria Anisimova

Institute of Applied Simulations, School of Life Sciences and Facility Management, Zürich University of Applied Sciences, CH-8820 Wädenswil, Switzerland

Type of article: Mini-review

Address: Maria Anisimova**,** Einsiedlerstrasse 31a**,** 8820 Wädenswil, Switzerland

Phone: +41 (0) 58 934 58 82

E-Mail: maria.anisimova@zhaw.ch

Key words: molecular evolution, applied bioinformatics, modeling, selection, adaptation, conservation, drug target, resistance, immune response

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.427v1 | CC-BY 4.0 Open Access | received: 28 Jun 2014, published: 28 Jun 2014

Abstract

Today computational molecular evolution and bioinformatics are vibrant research areas that flourish on large amounts of complex datasets generated by new generation technologies – from full genomes and proteomes to microbiomes, metabolomes and epigenomes. Yet the foundations for successful mining and the analyses of such data were established long before the structure of the DNA was discovered. Darwin's theory of evolution by means of natural selection not only remains relevant today, but also provides solid ground for computational research with a variety of applications. The data size and its complexity require empirical scientists to work in close collaboration with experts in computational science, modeling and statistics, as Sir R. Fisher has beautifully demonstrated in early 20th century. Particularly, modern computational methods for evaluating selection in molecular sequences are very useful for generating biological hypotheses and candidate gene sets for follow-up experiments. Evolutionary analyses of selective pressures in genomic data have high potential for applications, since natural selection is a leading force in function conservation, in adaptation to emerging pathogens, new environments, and plays key role in immune and resistance systems. At this stage, pharma and biotech industries can successfully use this potential, taking the initiative to enhance their research and development with the state-of the art bioinformatics approaches. This mini-review provides a quick "why-and-how" guide to the current approaches that apply the evolutionary principles of natural selection to real life problems – from drug target validation, vaccine design and protein engineering to applications in agriculture, ecology and conservation.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.427v1 | CC-BY 4.0 Open Access | received: 28 Jun 2014, published: 28 Jun 2014

The role of computational scientists in genomics and its applications

For well over a century computational scientists have been faithfully working side-by-side empirical life scientists, supporting key developments in molecular and evolutionary biology. While the impact of such work is easy to overlook, progress in life sciences heavily relies on solid backing from statisticians and computational scientists. One striking example is the monumental contribution of Sir Ronald A. Fisher, one of the first *bioinformaticians*. Fisher single-handedly developed the essential statistical theory for experimental design and hypothesis testing, contributing many widely used techniques such as the analysis of variance and the method of maximum likelihood – originally, to address the needs of the agricultural research at the Rothamstead Experimental Station in Harpenden, UK. Today many disciplines beyond life sciences rely on this methodology. Prominently, together with S. Wright and J. B. S. Haldane, Fisher has established the field of population genetics, becoming one of the founders of neo-Darwinian evolutionary synthesis – the current paradigm of the evolutionary biology, where the principles of Mendelian genetics were reconciled with Darwin's theory of evolution by natural selection at the level of hereditary molecular information.

The discovery of the hereditary DNA molecule and its structure was followed by rapid progress in sequencing technologies. Developments of statistical methodology kept pace: we now also have a wide range of excellent statistical methods for analyzing these vast genomics data so we can make inferences useful not only for our fundamental understanding of molecular evolution but also for applications in medical genetics, pharmacology, biotechnology, agriculture and ecology. The size and the complexity of molecular data underline the importance of the interdisciplinary collaborations and the crucial role of statisticians and computational scientists for the success of data exploration in projects using genomics and omics data. The level of biological data complexity has clearly passed the so-called “Excel barrier”, and the industry using genomics data can no longer rely on old practices. Consequently, pharma and biotechnology companies saw an increasing demand for professional computational scientists with strong skills in mathematical modeling, machine learning, data mining, complex optimization and data representation (e.g., Price 2012). Bioinformatics and computational biology are now embracing the challenges of translational research.

The importance of selection studies at the genomic level

The field of computational genomics has been growing steadily attracting more research funding for both academic and applied research in biotech and pharma companies. Here I focus on the potential of computational methods to study how genomic changes occur over time and their impact on phenotype or genetic fitness (Yang 2006, Anisimova 2012, Cannarozzi and Schneider 2012). While Darwin has described how selection may act on a phenotype, he had no knowledge of hereditary mechanisms, and would have been pleased to see how far we have come today in our understanding of selective mechanisms in molecular sequences. Current computational methods can detect genomic regions under selection and help to elaborate on the biological mechanisms generating the observed molecular patterns. Considering this, computational methods provide effective means of narrowing down the space of plausible candidates or hypotheses for further testing. A diversity of biological mechanisms may cause genetic mutations with various fitness effects, leading to a variety of ways natural selection can manifest itself. The central role of selective mechanisms at the molecular level has been demonstrated in the adaptation to new environments, the host-pathogen “arms” race, the emergence of competition, the evolution of complexity, and in the morphological and behavioral evolution, for example see fig. 1 of (Anisimova and Liberles 2012). Natural

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.427v1 | CC-BY 4.0 Open Access | received: 28 Jun 2014, published: 28 Jun 2014

selection may act on the protein, on the DNA sequence, and even on whole genomic features. Negative or purifying selection conserves the sequence (or other molecular features), while positive selection acts in a diversifying or a directional manner favoring specific changes. Positive selection typically affects molecular regions involved in genetic conflict, and often acts in an episodic manner – limited to certain time periods. Studies of selective pressures across a genome help to understand the biological constraints and to identify the mutational hotspots due to adaptive processes. This is why selection scans became an indispensible element of modern genomic studies (e.g., Stapley, Reger et al. 2010, Fu and Akey 2013).

Studies of selective constraints in genomic sequences from populations and species can have a variety of applications. Identification of deleterious mutations (e.g., mutations causing disease) may aid the development of gene therapies and personalized treatments. Detecting hotspots of diversifying pressure in antigenic sites, epitopes and pathogenic receptors can be used in drug and vaccine design. Phylogenetic methods are increasingly used in immunology and cancer genomics. The analysis of selective pressures and disease transmission rates using host and pathogen samples provides important clues for epidemiology, helping to understand the disease dynamics and to develop predictive strategies for disease control. This applies equally to animal and plant hosts as well as their pathogens, thus having applications also in the domain of agricultural research such as developing molecular-based strategies for increasing crop resistance to pathogens. Similarly, evolutionary studies may provide insights to the genetic basis for stress tolerance and yields of animal and plant products. Other applications of molecular evolution and selection analyses may include biodiversity, conservation, sustainable development, bioremediation, bioengineering and nutrition. Below I briefly draw attention to some successful approaches for studying the evolutionary dynamics in molecular sequences, illustrated by examples (summarized in Table 1).

Computational approaches for evaluating evolution and selection in molecular sequences

Evaluating selective pressures in molecular sequences relies on the comparative evolutionary approach, and therefore requires at least two homologous sequences (Nielsen and Hubisz 2005). The power of the approach depends on the number and the range of sequences analyzed (Anisimova, Bielawski et al. 2001). For large samples from well-designed experiments, it is possible to accurately predict the positions and the time episodes where selection has operated (Anisimova, Bielawski et al. 2002, Anisimova and Yang 2007, Lu and Guindon 2014). The basic idea behind all tests for selection is to compare the molecular patterns observed in genomic sequences to what could be expected by chance. Significant deviations point to interesting candidate regions, sites or time episodes, and provide excellent hypotheses for further experimental and statistical testing. Different methods use different statistics to make their inferences about selection. Stochastic modeling of molecular changes through time has been particularly successful, typically employing Markov models of character substitution. Among widely used methods are likelihood ratio tests of codon substitution models, which detect selection on the protein sequence using the comparison of nonsynonymous (amino-acid altering) and synonymous (amino-acid preserving) substitution rates (for review see Kosiol and Anisimova 2012). If a test is significant, Bayesian prediction is used to identify the selected positions or lineages affected by selection. The pharmaceutical giant GlaxoSmithKline (GSK) acknowledged the applied value of this methodology by an award to the principal investigator Prof Ziheng Yang (UCL, UK). The relevance of selection analyses using codon models for downstream applications can be demonstrated with a selection of case studies. A classic example is the human major histocompatibility complex molecules of class I (glycoproteins mediating cellular immunity against intracellular pathogens), where all residues under diversifying selection pressure were found clustered in the antigen recognition site (Hughes and Nei 1988, Yang and Swanson 2002). In another example, selection analyses identified a sequence region of 13 amino acids with many positive-selected sites in TRIM5α, involved in cellular antiviral defense (Sawyer, Wu et al. 2005). Functional studies of chimeric TRIM5α genes showed that the detected region was responsible for the difference in function between the rhesus monkey linage where TRIM5α restricts HIV-1 and the human TRIM5α that has only weak restriction.

More generally, the numerous genome-wide scans in mammals agree that genes affected by positive diversifying selection are largely responsible for sensory perception, immunity and defense functions (Kosiol, Vinar et al. 2008). Consequently, pharma and biotech companies should make a greater use of computational approaches to detect genes and biochemical pathways subject to differential adaptive evolution in human and other lineages used as experimental model organisms (Vamathevan, Hasan et al. 2008, Vamathevan, Hall et al. 2013). This type of studies can be valuable for example when selecting drug targets. Particularly, evolutionary analyses can pinpoint evolutionary differences between model organisms used for drug target selection. Such differences can be responsible for unpredicted disparities in response to medical treatment, as it has been highlighted by the tragic effects of TGN1412 treatment during human drug trials in 2006 (Stebbings, Poole et al. 2009). Selection analyses are also important for research in agriculture or conservation, since in plant genomes positive selection affects most notably disease resistance genes (Meyers, Shen et al. 1998, Mondragon-Palomino, Meyers et al. 2002), defense enzymes such as chitinases (Bishop, Dean et al. 2000) and genes responsible for stress tolerance (Roth and Liberles 2006). Consequently evolutionary studies help to detect proteins, binding sites and their interactions relevant for host-pathogen coevolution. For example, diversifying selective pressure drives the evolution of several exposed residues in leucine-rich repeats (LRRs) of the bacterial type III effectors (that attack plant defense system) from the phytopathogenic *R. Solancearum* infecting hundreds of plant varieties including agriculturally important crops (Kajava, Anisimova et al. 2008). Similarly, studies of phylogenetic diversity and selection in viral strains and antibody sequences are contributing to the new HIV vaccine development strategy, whereby antibodies are designed to bind to conserved epitopes of selected viral targets (de Oliveira, Salemi et al. 2004, Mouquet, Klein et al. 2011, Scheid, Mouquet et al. 2011, Klein, Mouquet et al. 2013). Moreover, molecular evolution modeling approaches can greatly enhance the modeling of antigenic dynamics of pathogens over time (e.g., Bedford, Suchard et al. 2014).

In protein coding sequences selection may also act on the DNA, whereby synonymous codon changes may affect protein's stability, expression, structure and function (Komar 2007, Plotkin and Kudla 2011). Translational selection manifests itself as the overall codon bias in a gene to match the abundances of cognate tRNA. Remarkably, this property can be successfully used in biotechnology, for example to dramatically increase transgene expression by synthesizing sequences with optimal synonymous codons (Gustafsson, Govindarajan et al. 2004). Optimal codon usage may be approximated by codon usage bias – using bioinformatics methods (Roth, Anisimova et al. 2012). Besides this, more subtle selective mechanisms may act on certain codon positions. These mechanisms may result in synonymous changes that can affect protein structure, abundance and function. In human genes this may lead to disease or may be responsible for differences in individual responses to drug treatment. Haplotypes with synonymous changes may have increased fitness and will be consequently increase in frequency in a population. Therefore, the knowledge of these specific synonymous polymorphisms may be important to explain differential treatment effects in population and contribute to the development of personalized medicines. Molecular evolution methods are powerful enough to detect such interesting candidate cases: Recent study of synonymous rates detected many disease related genes, particularly associated with various cancers, as well as many metabolizing enzymes and transporters, which affect the disposition, safety and efficacy of small molecule drugs in pharmacogenetics (Dimitrieva and Anisimova 2014). This shows that computational molecular evolution studies have real power to predict genes and codon positions where a replacement of synonymous codons changes protein fitness. Such predictions promise to be valuable for applications in protein engineering. Indeed, some biotech companies such as DAPCEL are already using the knowledge of interesting synonymous positions for enhanced protein production. Compared to laborious and time-consuming trial-and error experiments, computational prediction offers a fast way of obtaining candidate genes and positions for experimental validation. Furthermore, monitoring of the synonymous rates may be also informative for diagnostics purposes, as has been shown in evolutionary studies of serial viral samples from HIV-positive patients (Lemey, Kosakovsky Pond et al. 2007).

Species evolution is however a result of complex population dynamics, making population scale studies of genetic diversity a powerful complement to codon-based selection analyses. Successful population level techniques include tests of neutrality (Nielsen 2001), Poisson random-field models (e.g., Sawyer and Hartl 1992, Amei and Smith 2014) combined with demographic modeling and genome-wide association studies (Besenbacher, Mailund et al. 2012). These methods apply to full genome sequences helping to identify also non-coding genomic regions of functional relevance and those associated with certain population traits. For medical genetics, uncovering the relevance of genomic variation in populations helps to pinpoint the disease variants and use this information in the development of personalized medicines and treatments. Determining fitness of specific mutations is now possible using macro-evolutionary inferences and population genetics approaches (Boyko, Williamson et al. 2008, Chun and Fay 2009, Adzhubei, Schmidt et al. 2010), which can be successfully combined with genome-wide association studies (Manolio, Collins et al. 2009). These inferences could be combined with applications in a clinical context (Ashley, Butte et al. 2010).

However, many traits are shaped by multiple loci so that the effects of any single mutation can be observed only through their epistatic effects (Phillips 2008, Stranger, Stahl et al. 2011). Consequently, computational approaches recently extended single loci inferences to detecting epistatic effects of mutations through the identification of polygenic selection, i.e., whereby selection affects whole gene clusters whose protein products interconnected in the biological pathways that they share. Such analyses found that polygenic selection often affects pathways involved in immune response and adaptation to pathogens (Daub, Hofer et al. 2013), which is also consistent with results from single loci studies.

Another approach for detecting selective signatures is based on detecting shifts in evolutionary substitution rates over time, for example based on covarion or Markov modulated models (Galtier 2001, Guindon, Rodrigo et al. 2004). Such methods may be used to detect functional shifts in proteins of interest, providing evolutionary information that aids structural and functional protein prediction. Therefore such analyses can be helpful for many pharma and biotech applications that use structural modeling to design proteins and peptides for therapeutic or other biotechnology applications (e.g., Khoury, Smadbeck et al. 2014). Alternatively, changing diversification rates can provide evidence for changing environments, emerging pathogens and shed light on epidemiological dynamics. Diversification bursts or exponential growth, for example, may represent the emergence of particularly virulent strains resulting in epidemics. Such selective signatures can be characterized by the birth-death models describing on stochastic branching processes as phylogenies or genealogies relating molecular sequence samples (Stadler, Kuhnert et al. 2013). This approach allows to evaluate the effects of public health interventions by estimating the rates of transmission, recovery, and sampling, and consequently, the effective reproductive number. For epidemiology-related problems, these techniques become particularly powerful when combined with classical epidemiologic models SIR or SIS (Kuhnert, Stadler et al. 2014, Leventhal, Gunthard et al. 2014). Application of evolutionary methods may be also useful for the analyses of somatic hypermutation in antibody sequences during antibody maturation, or to monitor somatic mutations in cancerous tissues (Litman, Cannon et al. 2005, Campbell, Pleasance et al. 2008, Yates and Campbell 2012, Zhu, Ofek et al. 2013). Indeed, applications of phylogenetic methods to cancer and immunology research are now attracting more attention and funding.

Selection may also operate on whole genomic features, such as indels, gene order, gene copy numbers, transposable elements, miRNAs, post-translational modifications, etc. To detect selective signatures of conservation or adaptation, the observed genomic patterns are compared with a neutral expectation, i.e., patterns that can arise by chance alone. For example, Schaper, Gascuel et al. (2014) proposed evaluating phylogenetic patterns produced by tandem repeats in eukaryotic proteins with respect to human lineage, in order to identify interesting candidate genes that might be under diversifying

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.427v1 | CC-BY 4.0 Open Access | received: 28 Jun 2014, published: 28 Jun 2014

pressures. In plants a similar analysis strongly pointed to lineages where diversification (in terms of unit number and their order conservation) occurs in LRRs that are found in abundance in plant resistance genes (Schaper and Anisimova 2014). Such analyses allow for example to pinpoint the relevant genes and lineages where selection on tandem repeat units is due to adaptation to emerging pathogens or to changing environmental conditions. This opens the door to applications such as synthetically introducing identified gene variants into plant genomes to produce crops with improved resistance or better stress tolerance properties.

Even when detecting selection is not a focal part of the analyses, modeling its influence on genomic data is of utmost importance. Failing to do so may lead to biased and inaccurate inferences that could misguide follow-up experimental studies. However, modeling selection enhances the predictive power of methods that are used to study adaptive or antagonistic processes. A nice example is the recent predictive fitness model for influenza, which couples the fitness values and frequencies of strains with molecular evolution modeling on an influenza stain genealogy for haemaglutinin gene (Luksza and Lassig 2014). This approach uses observed viral samples taken from year to year to predict evolutionary flu dynamics in the coming year, which is practically relevant for selecting vaccine strains for the new flu season.

Conclusions and perspectives

In summary, the last decades have seen the development of solid computational methodology that can accurately detect selective signatures at the molecular level. The methodological advancement is an open-ended process and will continue in order to address the challenges from new large and complex datasets. Yet, the application of the existing state-of the art already provides powerful means for the rapid generation of viable hypotheses and interesting candidates cases for further experimental testing. Genomics and omics data provides immense opportunities for applications in industry, but these need to be developed through close collaborations between computational and empirical scientists, with a continued feedback loop. Computational predictions provide ground for setting up new experiments that will generate new data with new levels of complexity. These data are then again analyzed by computational scientists, in order to refine the predictions and generate new hypotheses for further experimental validation (Figure 1). Statistical expertise is necessary to design new experimental setup. Large pharma and biotech companies have already seized upon this potential and use computational molecular evolution approaches for their translational research. This includes drug target identification and validation, animal model selection, preclinical safety assessment, vaccine design, epidemics control and drug repositioning. These techniques are promising to become mainstream, strengthening the current position of translational research in industry. The translational value of computational molecular evolution is not limited to health and pharma industry, but also include a variety of other exciting applications – protein engineering, agriculture, environmental risk assessment, ecology, biodiversity and conservation. Now that it is cheap and quick to generate genomic data, but greater thought should be invested into experimental design, which will make statistical and computational inferences more informative and more accurate. This cannot be done without strong interdisciplinary partnerships. Indeed, bioinformatics has now become a vibrant and highly interdisciplinary area of research and the outlook for its future and its applications is very optimistic – “Bioinformatics alive and kicking” (Stein 2008).

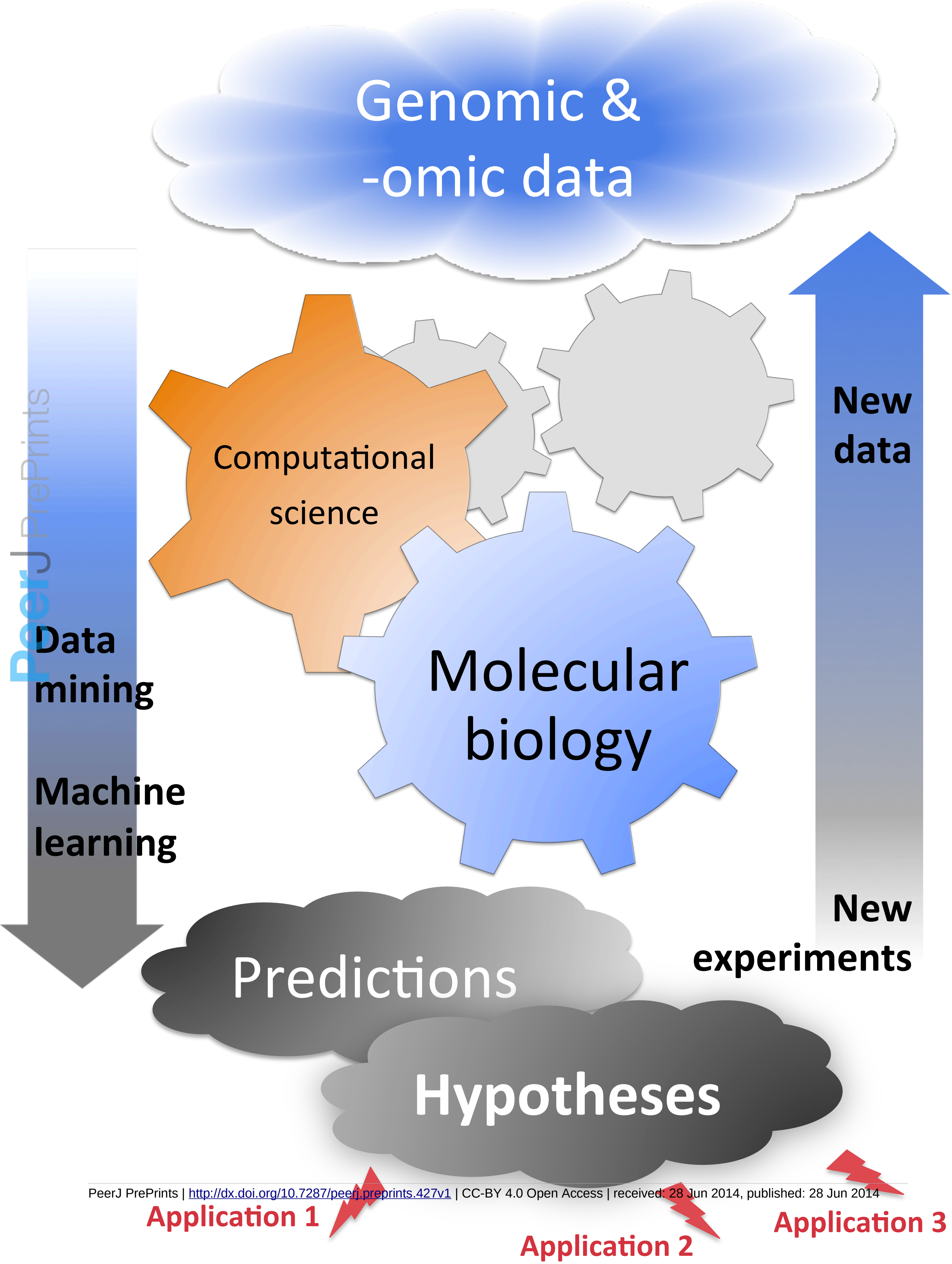
Acknowledgements

This article was motivated by the mini-symposium “Life in numbers: the use of modeling and simulation in life sciences” which took place on the 5th of June 2014 at the Zürich University of Applied Sciences (www.lifeinnumbers.ch) in partnership with the Swiss Institute of Bioinformatics and sponsored by Actelion, Philips and Geneious (Biomatters). The event intended to highlight the potential of computational biology in the applied domain and to bring together scientists from academia and industries. The first version of this article forms part of the Festschrift collection for the GNOME 2014 symposium to celebrate the retirement of Prof. Gaston Gonnet (ETH Zürich), and as such has been published open access as a pre-print with PeerJ.

Figure legends

Figure 1 – Feedback loop between experimental and computational stages of research and development. Applications of genomics and omics in industry arise through continuous collaborations between computational and empirical scientists, with a continued feedback loop: Computational predictions provide ground for setting up new experiments and generate new data with new levels of complexity. These data again analyzed by computational scientists to refine the predictions and generate new hypotheses for further experimental validation.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.427v1 | CC-BY 4.0 Open Access | received: 28 Jun 2014, published: 28 Jun 2014



PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.427v1 | CC-BY 4.0 Open Access | received: 28 Jun 2014, published: 28 Jun 2014

Table 1. Selected examples of applications of molecular evolution and selection studies

Application type	Description	Citation	Computational approach
Control of HIV infection	Protein function study of HIV restriction properties in TRIM5α	Sawyer, Wu et al. (2005)	Codon model tests for selection
Model species selection for pharmaceutical discovery	Assessment of pharmacological target homology	Vamathevan, Hall et al. (2013)	Phylogenetic analyses of gene families
HIV vaccine development	Assessment of phylogenetic diversity in viral proteins and antibodies; identification of conserved epitopes	de Oliveira, Salemi et al. (2004); Klein, Mouquet et al. (2013)	Phylogenetic analyses and codon model tests for selection
Flu epidemics prediction; vaccine strain selection	Modeling of antigenic dynamics of flu over time	Bedford, Suchard et al. (2014)	Phylogenetic diffusion model of antigenic evolution
Prediction of HIV progression	Monitoring the synonymous substitution rates in viral protein samples from HIV-positive patients over time	Lemey, Kosakovsky Pond et al. (2007)	"Relaxed-clock" modeling of codon evolution
Evaluating epidemics dynamics and the effect of public health interventions	Estimating the rates of transmission, recovery, sampling, and the effective reproductive number	Stadler, Kuhnert et al. (2013); Kuhnert, Stadler et al. (2014); Leventhal, Gunthard et al. (2014)	Birth-death phylogenetic models
Flu epidemics prediction; vaccine strain selection	Modeling adaptive epitope changes and deleterious mutations outside the epitopes in flu from one year to the next	Luksza and Lassig (2014)	Molecular evolution modeling over viral genealogies
Crop resistance	Identifying the resistant	Lee, Jia et al. (2011)	Analyses of genetic diversity and

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.427v1 | CC-BY 4.0 Open Access | received: 28 Jun 2014, published: 28 Jun 2014

complex disease biology; development personalized medicine	of genomic diversification, associations with diseases, estimating fitness of mutations	Chun and Fay (2009)	constraints, genome-wide association studies
***Disease biology; identification of vaccine targets**	Population genomics of the sexually transmitted bacteria *Chlamydia trachomatis*	Joseph, Didelot et al. (2012)	Genome-wide evolutionary analyses of conservation by codon models and population genetics approaches
***Disease biology**	Adaptation in the cavity causing bacteria *Streptococcus mutans*	Cornejo, Lefebure et al. (2013)	Genome-wide evolutionary analyses of conservation and demography
***Conservation and biodiversity; climate change**	Evaluating hybridization of blue whale subspecies in southern hemisphere	(Attard, Beheregaray et al. 2012)	Population genetics analyses
***Impact of climate change**	Evaluating the interplay between global climate change, genetic diversity and species interactions and community structure	Pauls, Nowak et al. (2013)	Evaluation of intraspecific genetic diversity by population genetics approaches

* Highlighted in the 2013 editorial "Highlights in applied evolutionary biology" in the peer-reviewed journal "Evolutionary Applications".

Cited literature

Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov and S. R. Sunyaev (2010). "A method and server for predicting damaging missense mutations." Nat Methods **7**(4): 248-249.

Amei, A. and B. T. Smith (2014). "Robust estimates of divergence times and selection with a poisson random field model: a case study of comparative phylogeographic data." Genetics **196**(1): 225-233.

Anisimova, M., Ed. (2012). Evolutionary genomics: statistical and computational methods. Methods in Mol Biol. New York, Humana press, Springer.

Anisimova, M., J. P. Bielawski and Z. Yang (2001). "Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution." Mol Biol Evol **18**(8): 1585-1592.

Anisimova, M., J. P. Bielawski and Z. Yang (2002). "Accuracy and power of bayes prediction of amino acid sites under positive selection." Mol Biol Evol **19**(6): 950-958.

Anisimova, M. and D. Liberles (2012). Detecting and understanding natural selection. Codon Evolution: mechanisms and models. G. Cannarozzi and A. Schneider. Oxford, Oxford University Press.

Anisimova, M. and Z. Yang (2007). "Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites." Mol Biol Evol **24**(5): 1219-1228.

Ashley, E. A., A. J. Butte, M. T. Wheeler, R. Chen, T. E. Klein, F. E. Dewey, J. T. Dudley, K. E. Ormond, A. Pavlovic, A. A. Morgan, D. Pushkarev, N. F. Neff, L. Hudgins, L. Gong, L. M. Hodges, D. S. Berlin, C. F. Thorn, K. Sangkuhl, J. M. Hebert, M. Woon, H. Sagreiya, R. Whaley, J. W. Knowles, M. F. Chou, J. V. Thakuria, A. M. Rosenbaum, A. W. Zaranek, G. M. Church, H. T. Greely, S. R. Quake and R. B. Altman (2010). "Clinical assessment incorporating a personal genome." The Lancet **375**(9725): 1525-1535.

Attard, C. R., L. B. Beheregaray, K. C. Jenner, P. C. Gill, M. N. Jenner, M. G. Morrice, K. M. Robertson and L. M. Moller (2012). "Hybridization of Southern Hemisphere blue whale subspecies and a sympatric area off Antarctica: impacts of whaling or climate change?" Mol Ecol **21**(23): 5715-5727.

Bedford, T., M. A. Suchard, P. Lemey, G. Dudas, V. Gregory, A. J. Hay, J. W. McCauley, C. A. Russell, D. J. Smith and A. Rambaut (2014). "Integrating influenza antigenic dynamics with molecular evolution." Elife **3**: e01914.

Besenbacher, S., T. Mailund and M. H. Schierup (2012). "Association mapping and disease: evolutionary perspectives." Methods Mol Biol **856**: 275-291.

Bishop, J. G., A. M. Dean and T. Mitchell-Olds (2000). "Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution." Proc Natl Acad Sci U S A **97**(10): 5322-5327.

Boyko, A. R., S. H. Williamson, A. R. Indap, J. D. Degenhardt, R. D. Hernandez, K. E. Lohmueller, M. D. Adams, S. Schmidt, J. J. Sninsky, S. R. Sunyaev, T. J. White, R. Nielsen, A. G. Clark and C. D. Bustamante (2008). "Assessing the evolutionary impact of amino acid mutations in the human genome." PLoS Genet **4**(5): e1000083.

Campbell, P. J., E. D. Pleasance, P. J. Stephens, E. Dicks, R. Rance, I. Goodhead, G. A. Follows, A. R. Green, P. A. Futreal and M. R. Stratton (2008). "Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing." Proc Natl Acad Sci U S A **105**(35): 13081-13086.

Cannarozzi, G. M. and A. Schneider, Eds. (2012). Codon Evolution: Mechanisms and Models. UK, Oxford University Press.

Chun, S. and J. C. Fay (2009). "Identification of deleterious mutations within three human genomes." Genome Res **19**(9): 1553-1561.

Cornejo, O. E., T. Lefebure, P. D. Bitar, P. Lang, V. P. Richards, K. Eilertson, T. Do, D. Beighton, L. Zeng, S. J. Ahn, R. A. Burne, A. Siepel, C. D. Bustamante and M. J. Stanhope (2013). "Evolutionary and population genomics of the cavity causing bacteria Streptococcus mutans." Mol Biol Evol **30**(4): 881-893.

Daub, J. T., T. Hofer, E. Cutivet, I. Dupanloup, L. Quintana-Murci, M. Robinson-Rechavi and L. Excoffier (2013). "Evidence for polygenic adaptation to pathogens in the human genome." Mol Biol Evol **30**(7): 1544-1558.

de Oliveira, T., M. Salemi, M. Gordon, A. M. Vandamme, E. J. van Rensburg, S. Engelbrecht, H. M. Coovadia and S. Cassol (2004). "Mapping sites of positive selection and amino acid diversification in the HIV genome: an alternative approach to vaccine design?" Genetics **167**(3): 1047-1058.

Dimitrieva, S. and M. Anisimova (2014). "Unraveling Patterns of Site-to-Site Synonymous Rates Variation and Associated Gene Properties of Protein Domains and Families." PLoS One **9**(6): e95034.

Fu, W. and J. M. Akey (2013). "Selection and adaptation in the human genome." Annu Rev Genomics Hum Genet **14**: 467-489.

Galtier, N. (2001). "Maximum-likelihood phylogenetic analysis under a covarion-like model." Mol Biol Evol **18**(5): 866-873.

Guindon, S., A. G. Rodrigo, K. A. Dyer and J. P. Huelsenbeck (2004). "Modeling the site-specific variation of selection patterns along lineages." Proc Natl Acad Sci U S A **101**(35): 12957-12962.

Gustafsson, C., S. Govindarajan and J. Minshull (2004). "Codon bias and heterologous protein expression." Trends Biotechnol **22**(7): 346-353.

Hughes, A. L. and M. Nei (1988). "Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection." Nature **335**(6186): 167-170.

Joseph, S. J., X. Didelot, J. Rothschild, H. J. de Vries, S. A. Morre, T. D. Read and D. Dean (2012). "Population genomics of Chlamydia trachomatis: insights on drift, selection, recombination, and population structure." Mol Biol Evol **29**(12): 3933-3946.

Kajava, A. V., M. Anisimova and N. Peeters (2008). "Origin and Evolution of GALA-LRR, a New Member of the CC-LRR Subfamily: From Plants to Bacteria?" PLoS ONE **3**(2): e1694.

Khoury, G. A., J. Smadbeck, C. A. Kieslich and C. A. Floudas (2014). "Protein folding and de novo protein design for biotechnological applications." Trends Biotechnol **32**(2): 99-109.

Klein, F., H. Mouquet, P. Dosenovic, J. F. Scheid, L. Scharf and M. C. Nussenzweig (2013). "Antibodies in HIV-1 vaccine development and therapy." Science **341**(6151): 1199-1204.

Komar, A. A. (2007). "Genetics. SNPs, silent but not invisible." Science **315**(5811): 466-467.

Kosiol, C. and M. Anisimova (2012). "Selection on the protein-coding genome." Methods Mol Biol **856**: 113-140.

Kosiol, C., T. Vinar, R. R. da Fonseca, M. J. Hubisz, C. D. Bustamante, R. Nielsen and A. Siepel (2008). "Patterns of positive selection in six Mammalian genomes." PLoS Genet **4**(8): e1000144.

Kuhnert, D., T. Stadler, T. G. Vaughan and A. J. Drummond (2014). "Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death SIR model." J R Soc Interface **11**(94): 20131106.

Lee, S., Y. Jia, M. Jia, D. R. Gealy, K. M. Olsen and A. L. Caicedo (2011). "Molecular evolution of the rice blast resistance gene Pi-ta in invasive weedy rice in the USA." PLoS One **6**(10): e26260.

Lemey, P., S. L. Kosakovsky Pond, A. J. Drummond, O. G. Pybus, B. Shapiro, H. Barroso, N. Taveira and A. Rambaut (2007). "Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics." PLoS Comput Biol **3**(2): e29.

Leventhal, G. E., H. F. Gunthard, S. Bonhoeffer and T. Stadler (2014). "Using an epidemiological model for phylogenetic inference reveals density dependence in HIV transmission." Mol Biol Evol **31**(1): 6-17.

Litman, G. W., J. P. Cannon and L. J. Dishaw (2005). "Reconstructing immune phylogeny: new perspectives." Nat Rev Immunol **5**(11): 866-879.

Lu, A. and S. Guindon (2014). "Performance of standard and stochastic branch-site models for detecting positive selection among coding sequences." Mol Biol Evol **31**(2): 484-495.

Luksza, M. and M. Lassig (2014). "A predictive fitness model for influenza." Nature **507**(7490): 57-61.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll and P. M. Visscher (2009). "Finding the missing heritability of complex diseases." Nature **461**(7265): 747-753.

Meyers, B. C., K. A. Shen, P. Rohani, B. S. Gaut and R. W. Michelmore (1998). "Receptor-like genes in the major resistance locus of lettuce are subject to divergent selection." Plant Cell **10**(11): 1833-1846.

Mondragon-Palomino, M., B. C. Meyers, R. W. Michelmore and B. S. Gaut (2002). "Patterns of positive selection in the complete NBS-LRR gene family of Arabidopsis thaliana." Genome Res **12**(9): 1305-1315.

Mouquet, H., F. Klein, J. F. Scheid, M. Warncke, J. Pietzsch, T. Y. Oliveira, K. Velinzon, M. S. Seaman and M. C. Nussenzweig (2011). "Memory B cell antibodies to HIV-1 gp140 cloned from individuals infected with clade A and B viruses." PLoS One **6**(9): e24078.

Nielsen, R. (2001). "Statistical tests of selective neutrality in the age of genomics." Heredity **86**(Pt 6): 641-647.

Nielsen, R. and M. J. Hubisz (2005). "Evolutionary genomics: detecting selection needs comparative data." Nature **433**(7023): E6; discussion E7-8.

Pauls, S. U., C. Nowak, M. Balint and M. Pfenninger (2013). "The impact of global climate change on genetic diversity within populations and species." Mol Ecol **22**(4): 925-946.

Phillips, P. C. (2008). "Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems." Nat Rev Genet **9**(11): 855-867.

Plotkin, J. B. and G. Kudla (2011). "Synonymous but not the same: the causes and consequences of codon bias." Nat Rev Genet **12**(1): 32-42.

Price, M. (2012). "Computational Biologists: The Next Pharma Scientists?" Science Careers(April 13).

Roth, A., M. Anisimova and G. Cannarozzi (2012). Measuring codon-usage bias. Codon Evolution: mechanisms and models. G. Cannarozzi and A. Schneider. Oxford, Oxford University Press.

Roth, C. and D. A. Liberles (2006). "A systematic search for positive selection in higher plants (Embryophytes)." BMC Plant Biol **6**: 12.

Sawyer, S. A. and D. L. Hartl (1992). "Population genetics of polymorphism and divergence." Genetics **132**(4): 1161-1176.

Sawyer, S. L., L. I. Wu, M. Emerman and H. S. Malik (2005). "Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain." Proc Natl Acad Sci U S A **102**(8): 2832-2837.

Schaper, E. and M. Anisimova (2014). "The Evolution and Function of Protein Tandem Repeats in Plants." New Phytol(under review).

Schaper, E., O. Gascuel and M. Anisimova (2014). "Deep conservation of human protein tandem repeats within the eukaryotes." Mol Biol Evol **31**(5): 1132-1148.

Scheid, J. F., H. Mouquet, B. Ueberheide, R. Diskin, F. Klein, T. Y. Oliveira, J. Pietzsch, D. Fenyo, A. Abadir, K. Velinzon, A. Hurley, S. Myung, F. Boulad, P. Poignard, D. R. Burton, F. Pereyra, D. D. Ho, B. D. Walker, M. S. Seaman, P. J. Bjorkman, B. T. Chait and M. C. Nussenzweig (2011). "Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding." Science **333**(6049): 1633-1637.

Stadler, T., D. Kuhnert, S. Bonhoeffer and A. J. Drummond (2013). "Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV)." Proc Natl Acad Sci U S A **110**(1): 228-233.

Stapley, J., J. Reger, P. G. Feulner, C. Smadja, J. Galindo, R. Ekblom, C. Bennison, A. D. Ball, A. P. Beckerman and J. Slate (2010). "Adaptation genomics: the next generation." Trends Ecol Evol **25**(12): 705-712.

Stebbings, R., S. Poole and R. Thorpe (2009). "Safety of biologics, lessons learnt from TGN1412." Curr Opin Biotechnol **20**(6): 673-677.

Stein, L. D. (2008). "Bioinformatics: alive and kicking." Genome Biol **9**(12): 114.

Stranger, B. E., E. A. Stahl and T. Raj (2011). "Progress and promise of genome-wide association studies for human complex trait genetics." Genetics **187**(2): 367-383.

Vamathevan, J. J., M. D. Hall, S. Hasan, P. M. Woollard, M. Xu, Y. Yang, X. Li, X. Wang, S. Kenny, J. R. Brown, J. Huxley-Jones, J. Lyon, J. Haselden, J. Min and P. Sanseau (2013). "Minipig and beagle animal model genomes aid species selection in pharmaceutical discovery and development." Toxicol Appl Pharmacol **270**(2): 149-157.

Vamathevan, J. J., S. Hasan, R. D. Emes, H. Amrine-Madsen, D. Rajagopalan, S. D. Topp, V. Kumar, M. Word, M. D. Simmons, S. M. Foord, P. Sanseau, Z. Yang and J. D. Holbrook (2008). "The role of positive selection in determining the molecular cause of species differences in disease." BMC Evol Biol **8**: 273.

Yang, Z. (2006). Computational molecular evolution. Oxford, Oxford University Press.

Yang, Z. and W. J. Swanson (2002). "Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes." Mol Biol Evol **19**(1): 49-57.

Yates, L. R. and P. J. Campbell (2012). "Evolution of the cancer genome." Nat Rev Genet **13**(11): 795-806.

Zhu, J., G. Ofek, Y. Yang, B. Zhang, M. K. Louder, G. Lu, K. McKee, M. Pancera, J. Skinner, Z. Zhang, R. Parks, J. Eudailey, K. E. Lloyd, J. Blinn, S. M. Alam, B. F. Haynes, M. Simek, D. R. Burton, W. C. Koff, N. C. S. Program, J. C. Mullikin, J. R. Mascola, L. Shapiro and P. D. Kwong (2013). "Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains." Proc Natl Acad Sci U S A **110**(16): 6470-6475.