Funct Integr Genomics (2015) 15:121–129
DOI 10.1007/s10142-015-0434-3

REVIEW

Discoveries and advances in plant and animal genomics

Rudi Appels · Johan Nystrom · Hollie Webster · Gabriel Keeble-Gagnere

Received: 12 February 2015 / Revised: 17 February 2015 / Accepted: 19 February 2015 / Published online: 13 March 2015

Abstract Plant and animal genomics is a broad area of research with respect to the biological issues covered because it continues to deal with the structure and function of genetic material underpinning all organisms. This mini-review utilizes the plenary lectures from the Plant and Animal Genome Conference as a basis for summarizing the trends in the genome-level studies of organisms

Keyword Plant and animal genomes · ancient DNA · gene networks · databases · communication

Introduction

Plant and animal genomics is a diverse area with respect to the biological issues covered because it continues to deal with the structure and function of genetic material underpinning all organisms. The Plant and Animal Genome (PAG) Conference held in San Diego (California) in January each year provides an overview across all organisms at the genome level, and often it is evident that investment in the human area provides the leadership, applications and discoveries for researchers studying other organisms. This mini-review utilizes the plenary lectures from the conference as a basis for summarizing the trends in the genome-level studies of organisms, as indicated by the presentations from Phillip Bourne (National Institutes of Health, USA), Beth Shapiro (University of California Santa Cruz), Trey Ideker (University of California, San Diego), Xuemei Chen (University of California, Riverside), Mike Goddard (University of Melbourne, Victoria DEPI), Giles Oldroyd (John Innes Centre, UK) and Christina Warinner (University of Oklahoma). The research covered in this mini-review is based on published papers. Where unpublished information is cited, permission to include the information in this manuscript was obtained from the presenters.

R. Appels (✉) · J. Nystrom · H. Webster · G. Keeble-Gagnere
School of Veterinary and Life Sciences, Murdoch University, 90 South Street, Murdoch, Perth, Australia 6150
e-mail: r.appels@murdoch.edu.au

Evolutionary genomics

Species are the fundamental units in the biological classification of organisms and represent groups of organisms that interbreed and produce fertile progeny. In the presentation by Beth Shapiro, a detailed genome-level analysis of allele frequencies in some defined populations of polar and brown bears demonstrated that brown bears possess polar bear gene alleles across significant portions of their genomes, whereas brown bear gene alleles appeared to be absent from the polar bear genome (Cahill et al. 2013, 2015). The study was based on sequencing total DNA using either the 100 or 150-bp paired-end chemistry and Illumina HiSeq 2000, assembled using the polar bear genome sequence as a reference (Li et al. 2011). The distribution of allelic variation in both mitochondrial and nuclear DNA was consistent with gene dispersal via male brown bears. The asymmetry in gene allele distribution suggested that hybrid individuals were unable to backcross with the polar bear population. At a broader level, the asymmetric gene allele distribution in admixtures of closely related species has also been found in studies of the hybrid zones between *Mus domesticus* and *Mus musculus* (Teeter et al. 2008). The asymmetry was of a quantitative nature, possibly from asymmetric mating preferences between *M. domesticus* and *M. musculus*, resulting from urine signals recognized between *M. musculus* individuals, but not between *M. domesticus* individuals. The detailed study of Darwin's finches (*Geospiza*

fortis, *Geospiza scandens*, *Geospiza fuliginosa* and *Geospiza magnirostris*) also showed asymmetrical gene flow, which was dynamic and modified by environmental conditions (Grant and Grant 2010).

The asymmetrical gene flows were argued to provide a template for refining the analysis of interactions between climate, ecology and speciation. The plenary lecture by Christina Warinner discussed what is considered to be one of the clearest examples of gene flow influenced by culture/environment in human evolution, namely a genetic adaptation in the regulation of the lactase gene required for the digestion of milk lactose (Burger et al. 2007; Warinner et al. 2014a; Kruttli et al. 2014). The technology used in the studies described by Warinner focused on the ancient DNA and protein from the calculus on teeth of human skeletal remains (Warinner et al. 2014b). The amount of DNA extracted from calculus can be 3 orders of magnitude greater than that obtained from bone or dentine (Adler et al. 2013; Warinner et al. 2014a, b, c), and the tandem mass spectrometric-based analyses for protein has allowed fragments of protein to be interpreted, according to the deduced amino acid sequences of peptides (Warinner et al. 2014a). The technology developments were integral to the expanding field of palaeomicrobiology across a broad range of disciplines. In addition to the analysis of human archaeological faecal and bone/dental samples, the technologies included characterization of oil deposits, and permafrost and deep sea samples (Warinner 2014b). In the case of the human archaeological dental samples, whole genome sequencing of calculus utilized random primers to amplify and sequence a largely unbiased subset of the total DNA. For microbial DNA, the 16S ribosomal RNA (rRNA) amplicons recovered using short primers still provided a valuable source of DNA sequence for classifying the bacteria in the DNA sample under study (Warinner 2014d; current status of genome sequencing in bacteria generally is reviewed in Land et al. 2015, this issue). The analysis of the metagenomic database to assign ancient DNA sequences to bacterial classes and species was still a significant challenge, and it has been noted that the analysis of human gut bacteria using shotgun metagenomic approaches yielded lower diversity estimates than those based on amplicon sequencing of the 16S rRNA gene (Qin et al. 2010). The use of LC-mass spectrometry-based metaproteomics of the protein samples extracted from archaeological specimens, and digested with trypsin, was noted as becoming feasible (Cappellini et al. 2014; van Doorn et al. 2012) because reference datasets for interpreting the peptide sequences were being developed.

The analysis of both dental DNA and protein from human archaeological specimens provided the basis for investigating genetic adaptation in the human population in relation to the regulation of the lactase gene (Kruttli et al. 2014). The lactase enzyme hydrolyzes the lactose in milk to its component monosaccharides glucose and galactose, and persistence of its expression ('lactase persistence', LP) in adults is associated with the capacity to consume milk in European, African and Middle Eastern populations without unwanted side effects. The C-to-T single nucleotide polymorphism (SNP), (T-13910) located ca. 14,000 bp upstream of the lactase-phlorizin hydrolase (LCT) gene is associated with an absence of down-regulation of lactase activity after weaning. Assaying the T-13910 SNP in the ancient DNA samples suggested that the increased frequency of this mutation in Europe occurred in the period of 3000 BC to 1200 AD (Kruttli et al. 2014) since it appeared to be missing from the early Neolithic farmers in Europe (Burger et al. 2007). The increased frequency in the T-13910 SNP in the agricultural era in question was consistent with the fact that milk would have provided a clean, versatile and nutritious source of liquid (Lee et al. 1978) in early societies. Independent evidence from the direct assay of the milk whey protein beta-lactoglobulin in archaeological dental calculus in specimens dating back to at least the Bronze Age (ca. 3000 BC) in Europe and northern Southwest Asia (Warinner et al. 2014a) was also consistent with the timing of T-13910 frequency increases. The analysis of dental calculus was thus providing clear indicators for dietary variables driving recent natural selection in humans (Adler et al. 2013; Warriner 2014b).

Advances in agriculture sciences

The plenary lecture by Xuemei Chen provided new insights into the control of microRNAs with a particular reference to flowering in plants using *Arabidopsis* as a model. It is generally acknowledged that small RNAs (20–24 nucleotides) are involved in the control of gene expression in a wide range of gene networks and that their steady-state levels need to be carefully controlled (Ramachandran and Chen 2008). The family of miR172 microRNAs comprises five genes (MIR172a–e), and the respective promoters bind RNA polymerase II for producing the RNA products that repress the translation of the target *APETALA2* gene transcript (AP2; Chen 2004). The AP2 transcription factor contributes to the gene network that controls the developmental changes from growing stem meristem to forming flowers (Reinhardt and Kuhlemeier 2002; Chen 2004). Reduced levels of AP2 within the gene network are required for this developmental switch. The POWERDRESS (PWR) mutation was found to enhance the expression of MIR172a, MIR172b and MIR172c (leading to reduced AP2 expression) and hence defined one component of the extensive gene network controlling floral determinacy in the stem meristem (Yumul et al. 2013). The expression of MIR172d was independent of PWD. The existence of a family of microRNAs appeared to provide a well-buffered situation to ensure that AP2 protein levels were reduced as required for floral development.

In addition to transcription by RNA polymerase II, the in vivo levels of MIR172 microRNA were also determined by the activity of SMALL RNA DEGRADING NUCLEASE (SDN) genes, in particular SDN3 as judged from the analysis of mutations in the SDN genes (Ramachandran and Chen 2008). The factors controlling the levels of microRNAs more broadly were evidently the result of another network of genes determining the accessibility of microRNAs to 3′-exonuclease activity of SDN protein (Shen and Goodman 2004; Ren et al. 2014). The addition of U residues to the 5′ terminus of microRNA (miRNA) by miRNA nucleotidyl transferase (HESO1 in *Arabidopsis*) triggers their degradation in a process that is closely linked to the binding of miRNA to the argonaute-1 (AGO1) protein, the effector protein implementing the biological properties of miRNA (Ren et al. 2014). Methylation at the 2′-O position of miRNA by the HEN1 protein (in *Arabidopsis*; Yu et al. 2005) protects miRNA from degradation and uridylation (Ren et al. 2014), as judged from the analysis of mutations that suppress HEN1 activity. The wide influence of the regulation of miRNA has been highlighted in the analysis of tetraploid *Arabidopsis* lines (Wang et al. 2006; Ha et al. 2009) where hybrid lines between *Arabidopsis thaliana* and *Arabidopsis arenosa* showed non-additive expression of miRNAs and thus provide one mechanism for the expression of novel combination of genes in polyploids. Non-additive expression of gene per se has been reported in polyploids such as cotton and wheat (reviewed in Appels 2009), and the observations on the networks in play in *Arabidopsis* impact on the analysis of more complex genomes.

The status of genomics in *Arabidopsis* and rice is the most advanced among plants, and in many plant genomes, issues related to long-distance linkage and orientation of genome sequences remain to be a challenge. The new optical mapping technology (BioNano; Cao et al. 2014) has provided significant highlights across a broad range of organisms, and its application to the analysis of a complex genome such as wheat is discussed here as an illustration of the impact of this technology. At the genome sequence level, the International Wheat Genome Sequencing Consortium (IWGSC, http://www.wheatgenome.org/) has tackled the delineation of the genome structure using the capacity to generate bacterial artificial chromosome (BAC) libraries from flow-sorted chromosome arms followed by sequencing of BACs that were arranged into minimum tiling paths (Feuillet et al. 2012). This approach has been complemented by whole genome sequencing of DNA from flow-sorted chromosome arms (IWGSC 2014) as well as DNA from the entire genome (Brenchley et al. 2012; Chapman et al. 2015). A key resource for the assembly of genomes is a detailed reference genetic map since this relates genome sequence data back to the native DNA that exists in a nucleus of an organism (Feuillet et al. 2012). In this context, the optical mapping technologies are particularly significant because they are contributing to bridging the DNA sequence-to-whole chromosome gap (reviewed in Appels et al. 2014). The well-studied human genome has documented numerous structural variations, including 353,126 copy number variations and 1645 inversions, many of which are known to be associated with a wide range of medical issues in patients (Database of Genome Variants, http://dgv.tcag.ca/dgv/app/home). Optical mapping using nano-channels to array molecules of native DNA, plus advanced bioinformatics analyses, has been shown to provide a comprehensive assay system for defining a structural variation in DNA (Cao et al. 2014; O'Bleness et al. 2014). In the analysis by Cao et al. (2014), one labelled Nt.BspQI (nicking endonuclease) site was assayed, on average, every 9 kb in DNA molecules 1 Mb in length. A total of 932,855 molecules larger than 150 kb (223 Gb, ca. 70× average coverage) were studied. In aligning their optical maps to a reference genome sequence, Cao et al. (2014) estimated that the missing label rate was 10 % and the extra labelling rate was 17 %. Comparative analyses showed that the variation in complex regions such as the MHC loci could be readily assayed, in addition to establishing the number of repetitive sequences, INDELS and inversions at biologically significant loci. The optical mapping technology was a key tool in resolving errors in the 1q21.1–q21.2 region of human chromosome 1 (O'Bleness et al. 2014), particularly in the section of the genome carrying the repetitive Neuroblastoma Breakpoint Family (NBPF) genes.

The multiple, highly duplicated and complex regions in the human genome that remain largely intractable to analysis with commonly used assembly techniques are also present in other organisms. The application of the BioNano-based optical mapping to the D genome donor of hexaploid wheat (*Aegilops tauschii*) was presented in one of the IWGSC workshops by Mingcheng Luo (UC Davis). The sequence scaffolds (M Luo unpublished, based on Luo et al. 2013) were aligned to the optical maps, and this allowed scaffolds to be ordered and orientated into arrays of several megabases in length (Fig. 1a). Importantly, errors during assembly and scaffolding could be detected and resolved (Fig. 1b).

The genome sequence and annotation in cattle are well advanced, and the plenary lecture by Mike Goddard provided some benchmarks for the analysis of complex traits in cattle, as a model for advances in other organisms including humans and crops. Although the cattle reference genome is still a work in progress, many associations for complex traits such as carcass weight and milk production have been mapped into the high-density SNP maps available for cattle (Saatchi et al. 2014; Goddard 2014) in combination with pedigree studies (Haile-Mariam et al. 2013). The genome-wide association studies (GWASs) carried out on large cattle herds using the 50-K SNP chip (Saatchi et al. 2014) or whole genome sequencing (Daetwyler et al. 2014) has identified many

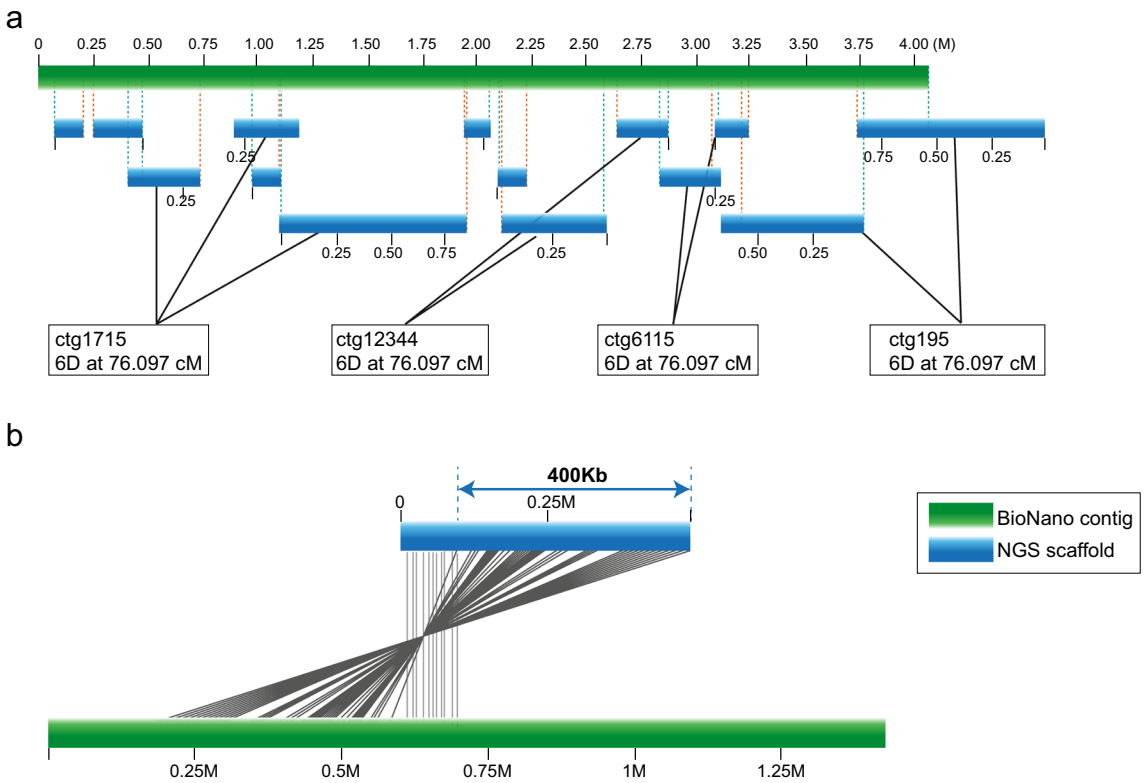

Fig. 1 Alignment of sequence scaffolds onto a BioNano contig. Scaffolds derived from MTP BAC clone sequences (Luo et al. 2013) were ordered and orientated onto a BioNano 4-Mb contig spanning (**a**). An example of an error during sequence scaffolding where approximately 400-kb sequence was placed on the opposite side (of correct side) in inverted orientation (*lower panel of figure*), which will ultimately guide editing sequence assembly

relatively small-effect associations rather than large-effect ones due to rare alleles (Goddard 2014). This has also been found in human GWAS analyses for the complex trait height (Yang et al. 2010, 2011). Increasing the density of SNPs from 50 to 800 K was not found to greatly increase the accuracy of phenotype prediction, but improvements in the analysis using BayesR were found to be useful (Erbe et al. 2012). At a broader level, the comparison of cattle breeds using the 50-K SNP identified regions of extended haplotype homozygosity (EHH) and these regions provided candidate genes for traits that characterize the respective breeds, locating to known QTL for the traits within breeds (Rothammer et al. 2013). The regions of EHH with candidate genes for traits were referred to as selection signatures and could cover up to 12-Mb sections of the genome sequence, but these also included 'short signatures' that housed, for example, the MAP2K6 gene that is associated with carcass weight, back fat and marbling in Korean beef cattle (Rothammer et al. 2013). A validation process identified known genes from independent studies, MSTN, within a selection signature region for double muscling in Blanc-Bleu Belge cattle and the gene MC1R within a red coat selection signature region for Red Holstein cattle. Many selection signatures did not provide annotated genes, and this is consistent with the observations that SNP association studies identified small-effect associations with complex traits. However, Rothammer et al. (2013) noted that 97 out of 480 of these signatures with poorly annotated genes overlapped between at least two breeds and could identify regions of particular interest for refining association studies.

The molecular genetic analysis of complex traits in cattle contrasted to the targeted study of nitrogen fixation in crops by Giles Oldroyd in his plenary lecture. The extensive knowledge base in the area of nitrogen fixation in legumes (Oldroy et al. 2011) has provided the basis for new investment in establishing this capacity in crops such as maize, with the view to then transferring to other non-legume crops (https://www.jic.ac.uk/news/2012/07/cereals-self-fertilise/#). It was argued that even the ability to sense nitrogen-fixing soil bacteria and develop simple associations with the roots could be very beneficial to the crop if no other sources of nitrogen were available. The legumes produce specialized organs, nodules, to optimize the symbiotic partnership between plant roots and rhizobia bacteria, but these refined structures were not necessarily a required outcome for nitrogen-fixing bacteria associated with roots to contribute to the supply of nitrogen to cereals.

In legumes, the early stages of nodule formation and bacterial infection require the recognition of the rhizobial

signalling molecule, a lipo-chitooligosaccharide, by receptor-like kinases (Oldroyd et al. 2011; Xie et al. 2012). The resulting fluctuation in calcium levels near the nuclei of root hair epidermal cells engages a calcium and calmodulin-dependent protein kinase (CCaMK) that phosphorylates a protein called CYCLOPS. The fluctuation in calcium levels is a characteristic feature of the recognition process and requires the symbiosis receptor-like kinase (SYMRK), components of nuclear pores and two cation channels (SYM pathway; Oldroyd et al. 2011). The next steps in the nodulation process require several transcription factors (NSP1, NSP2, ERN, NIN) to regulate gene expression for nodule formation to begin. The feasibility of establishing these early stages of the nitrogen fixation process in cereals has become evident from the study of arbuscular mycorrhiza (AM) where fungi form branched hyphal structures that penetrate root cortical cell to form a symbiotic relationship that promotes the exchange of nutrients (Brundrett 2002; Kistner and Parniske 2002; Yang et al. 2012). The majority of land plants form AM with zygomycete fungi (order Glomales; Brundrett 2002). Mutation studies in *Lotus japonica* have indicated that at least seven genetic loci control the early stages of establishing both AMs and nodulation (Kistner et al. 2005). Consistent with this finding, Chen et al. (2007, 2008) have shown that both CCaMK and CYCLOPS were required to establish AM in rice and that the rice CCaMK could complement mutations at the DMI3 locus in *Medicago truncatula* to re-establish nodulation. The studies by Gutjahr et al. (2008) showed that the SYM pathway was conserved by selecting rice lines with insertions (mutations) into analogs of the genes coding for CASTOR and POLLUX (required for initiating the Ca-spiking response) as well as insertions into genes coding for CCAMK and CYCLOPS. The analyses of the mutants validated the conclusions that AM colonization of rice roots utilized the same molecular machinery as used to form the early stage of rhizobia nodules.

A point of difference between the early stages of association to form either AM or nodules was the chemical nature of the signalling, namely strigolactone in the case of AM (Bonfante and Requena 2011) and lipo-chitooligosaccharides for rhizobia (Oldroyd et al. 2011; Rival et al. 2012; Chen et al. 2007, 2008). This difference would require a modification of the receptor in cereals in order to recognize the lipo-chitooligosaccharide for attracting the rhizobia and thus utilize some of the components of the molecular machinery for the nitrogen-fixing process that already exist in cereal crops.

Database analyses

The need to obtain a better value from the large investments in data generation from biological systems was discussed by Phillip Bourne. Laboratory research comprises multiple activities including experimental design, dataset construction, hypothesis formation and hypothesis testing. Formerly, these were usually carried out as integrated activities, but the era of data science is beginning to separate some of these activities (Bourne 2005) and dataset construction can now be carried out in its own right, in an open-ended way. While data publishers often have initial research problems in mind for their data, they might also hope that their data will ultimately be used to study entirely different, unforeseen problems. For example, datasets are now published independently in journals such as *GigaScience* (for example, Craft et al. 2014) and are then referenced in research articles based on the data. This open-endedness addresses the challenge of getting a better value out of large investments in data generation as part of what can be considered to be a digital enterprise (Bourne 2013). A component of this digital enterprise is the improvement and training in accessing methodology in order to work towards a better-shared understanding of how datasets should be produced and consumed. From a technical perspective, ontologies, such as gene ontology (GO, http://geneontology.org/) and EnvO (http://environmentontology.org/), and ontology frameworks, such as Resource Description Framework (RDF, http://www.w3.org/RDF/) and Web Ontology Language (OWL, http://www.w3.org/TR/owl-ref/), go a long way towards enabling the provision of context-independent, widely usable metadata. The adoption of ontology frameworks is a slow but ongoing process in the wider biology research community.

When applying existing datasets to new problems, data users may not always be able to communicate sufficiently with data generators and publishers. In order to maximize the reusability of datasets with limited communication, it is crucial that the datasets are accompanied by well-formulated metadata (Bourne 2005). Good metadata allows future users of a dataset to understand how it was created and how to integrate it with other processes and datasets in a scientifically sound approach. In this way, datasets have maximal mobility and value in their own right, following its separation from the environment in which they were generated.

At the highest and most general level of data integration, the data commons framework provides a conceptual basis for sharing, finding, integrating, reusing and attributing data. Central here is the need to assign a unique identifier (UID) to each dataset, for example a digital object identifier (DOI, http://www.doi.org/). The platform agnostic nature of the data commons framework lets it be deployed in a variety of environments, including a range of cloud platforms (Bourne 2013).

The capacity to obtain a better value from datasets in the area of associating gene networks with particular biological phenotypes was the focus in Trey Ideker's plenary lecture. Cytoscape (http://www.cytoscape.org/) was discussed as an environment for describing possible gene networks that may

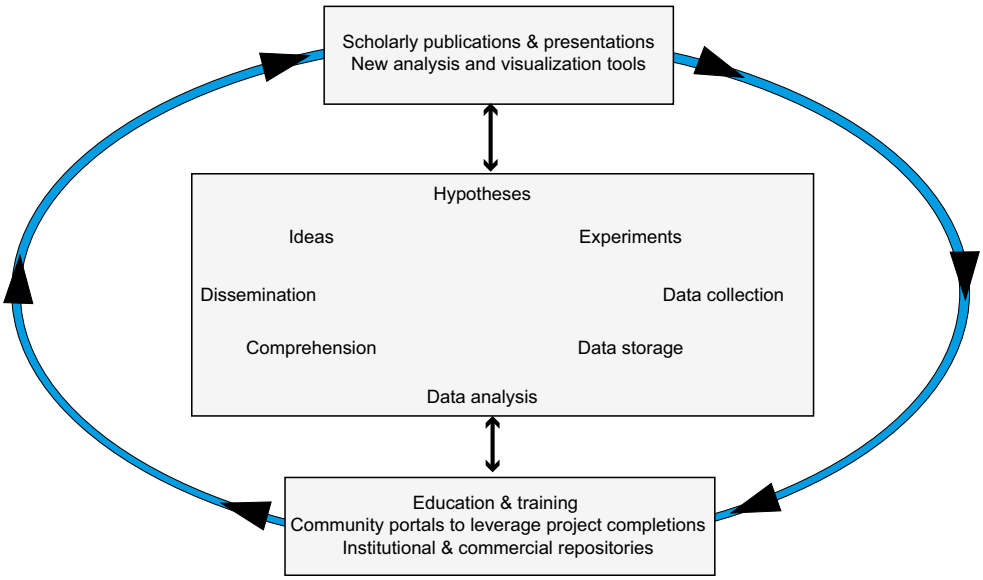

Fig. 2 A virtuous cycle based on Bourne (2013) to facilitate defining key aspects of the research cycle for making scientific research activity more sustainable

exist within the large datasets describing interactions between protein-protein, protein-DNA, kinase-substrate and gene co-expression at the transcriptome level (Saito et al. 2012). The community-based aspect of Cytoscape was facilitated by 152 publically available plug-ins validated by the Cytoscape developers (Saito et al. 2012). The capacity for Cytoscape to utilize gene network information to initiate the analysis of a complex dataset rather than having the gene network as an endpoint was illustrated in the study of several human diseases. The analysis of genotype to phenotype via networks and modules has been reviewed in Carter et al. (2013). Although biological networks described to date, especially in plants, fall short of capturing many important aspects of biological systems, the focus on using available networks and protein-protein interaction (PPI) information more effectively to interpret complex phenotypes has already been proven to be valuable. The network-based stratification (NBS)-based analysis of cancers (Hofree et al. 2013) provided a good example. The extensive database of mutations from the whole-of-genome analysis of tumours determined in The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) programs provided the starting point for the analysis. The variation in genome structure such as copy number variation, INDELS in genes and levels of transcription of genes allowed regions of gene networks to be defined that were characteristic of cancer subtypes (Hofree et al. 2013). The projection of genome-scale somatic mutation profiles onto gene interaction networks was built on prior developments in network-based prioritization of disease gene-protein complex associations by Vanunu et al. (2010). The NBS also utilized the consensus clustering software based on resampling (Monti et al. 2003), combining the outputs from 1000 subsamples, to establish clusters of higher confidence. The analysis established a valuable classification diagnostic for cancers.

In the analysis of the complex hereditary phenotype of neurodegenerative motor neuron diseases (hereditary spastic paraplegias (HSPs)) characterized by progressive age-dependent loss of corticospinal motor tract function, Novarino et al. (2014) used whole exome sequencing to identify 18 new candidate HSP genes. The extensive family pedigree analysis of mutations showed that *ERLIN1*, *KIF1C* and *NT5C2* were significant in the HSP phenotype (Novarino et al. 2014). In cases of family-specific mutations, supporting data from zebrafish functional (knock-down) experiments for genes such as *MARS* was obtained. At the network analysis level, the authors developed a HSP interactome network and this indicated that an apparently functionally diverse set of genes were, in fact, closely connected via basic biological processes, involving, for example, the endoplasmic reticulum. The interconnection of networks within the cell has been argued to form modules as part of the hierarchy of distributing function within a cell (Carter et al. 2013; Dutkowski et al. 2013; Mitra et al. 2013; Carvunis and Ideker 2014; Kramer et al. 2014).

Communication challenges

While essential to research, high-quality datasets are, by themselves, often not sufficient and unlocking the research potential of a particular dataset can sometimes require more sophisticated tools to provide helpful perspectives for lowering the barrier to exploration. As noted above, tools can provide capabilities for visualization, data integration from multiple sources, clustering and prioritization. Phillip Bourne

discussed a number of aspects of communication in science in the context of a virtuous cycle which, in part, is shown in Fig. 2 (modified from Bourne 2013) and highlights the key role of data collection, storage and analysis. The sustainability of research investment was argued to rely on the capacity to utilize large datasets in community collaboration, policy and infrastructure development as well as research and training.

The changing role of publications as the primary source of new information was argued to require re-evaluation in light of a prominent role of data collection, data storage and data analysis in the cycle shown in Fig. 2. The investment by funding agencies is usually viewed, by the administrators, as developing datasets and new solutions to issues such as arresting the growth of cancers and releasing crop varieties that are highly yielding when grown in dry environments. The importance of publishing in high impact factor journals was not necessarily rated as significant with the funding agencies as it did with the academics involved in the research.

A good meeting point between the differing views of research funders and research providers is to have the investment in data generation per se being more cost-effective. A key part of this cost-effectiveness was the need to improve the communication that can drive the utilization of large investments in data generation in multiple research projects. An important aspect of this was that high impact journals should take a greater role in ensuring compliance to the uploading of datasets into the digital environments (Alsheikh-Ali et al. 2011; Bourne 2013).

Acknowledgments The authors are grateful to Dr. Mingcheng Luo (UC Davis) for allowing his unpublished data to be cited. The authors are supported by core funding from Murdoch University and GRDC grant UMU00037.

Conflict of interest The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the reported research.

Ethical statement The manuscript is a review, and the authors have reported research based on published information unless otherwise acknowledged.

References

Adler CJ, Dobney K, Weyrich LS, Kaidonis J, Walker AW, Haak W, Bradshaw CJ, Townsend G, Soltysiak A, Alt KW, Parkhill J, Cooper A (2013) Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. Nat Genet 45:450e455

Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JPA (2011) Public availability of published research data in high-impact journals. PLoS One 6(9):e24357. doi:10.1371/journal.pone.0024357

Appels R (2009) Diversity of genome research at the 2009 Plant and Animal Genome Conference. Funct Integr Genomics 9(1):6. doi:10.1007/s10142-009-0112-4

Appels R, Nystrom-Persson J, Keeble-Gagnere G (2014) Advances in genome studies in plants and animals. Funct Integr Genomics 14:1–9. doi:10.1007/s10142-014-0364-5

Bonfante P, Requena N (2011) Dating in the dark: how roots respond to fungal signals to establish arbuscular mycorrhizal symbiosis. Curr Opin Plant Biol 14:451–457. doi:10.1016/j.pbi.2011.03.014

Bourne P (2005) Will a biological database be different from a biological journal? PLoS Comput Biol 1(3):e34

Bourne P (2013) CBIIT October 30, 2013. Bourne_slides_508_compliant

Brenchley R, Spannagl M, Pfeifer M, Barker GL, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, Kay S, Waite D, Trick M, Bancroft I, Gu Y, Huo N, Luo MC, Sehgal S, Gill B, Kianian S, Anderson O, Kersey P, Dvorak J, McCombie WR, Hall A, Mayer KF, Edwards KJ, Bevan MW, Hall N (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491:705–710. doi:10.1038/nature11650

Brundrett M (2002) Coevolution of roots and mycorrhizas of land plants. New Phytol 154:275–304

Burger J, Kirchner M, Bramanti B, Haak W, Thomas MG (2007) Absence of the lactase-persistence-associated allele in early Neolithic Europeans. P Natl Acad Sci USA 104:3736–3741

Cahill JA, Green RE, Fulton TL, Stiller M, Jay F et al (2013) Genomic evidence for island population conversion resolves conflicting theories of polar bear evolution. PLoS Genet 9(3):e1003345. doi:10.1371/journal.pgen.1003345

Cahill JA, Stirling I, Kistler L, Salamzade R, Ersmark E, Fulton TL, Stiller M, Green RE, Shapiro B (2015) Genomic evidence of geographically widespread effect of gene flow from polar bears into brown bears. Mol Ecol. doi:10.1111/mec.13038

Cao H, Hastie AR, Cao D, Lam ET, Su Y, Huang H, Liu X, Lin L, Andrews W, Chan S, Huang S, Tong X, Requa M, Anantharaman T, Krogh A, Yang H, Cao H, Xu X (2014) Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. GigaScience 3:34

Cappellini E, Collins MJ, Gilbert MTP (2014) Unlocking ancient protein palimpsests. Science 343:1320–1322. doi:10.1126/science.1249274

Carter H, Hofree M, Ideker T (2013) Genotype to phenotype via network analysis. Curr Opin Genet Dev 23:611–621. doi:10.1016/j.gde.2013.10.003

Carvunis AR, Ideker T (2014) Siri of the cell: what biology could learn from the iPhone. Cell 157:534–538. doi:10.1016/j.cell.2014.03.009

Chapman JA, Mascher M, Buluç AN, Barry K, Georganas E, Session A, Strnadova V, Jenkins J, Sehgal S, Oliker L, Schmutz J, Yelick KA, Scholz U, Waugh R, Poland JA, Muehlbauer GJ, Stein N, Rokhsar DS (2015) A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. Genome Biol 16(1):26

Chen X (2004) A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development. Science 303:2022–2025

Chen C, Gao M, Liu J, Zhu H (2007) Fungal symbiosis in rice requires an ortholog of a legume common symbiosis gene encoding a Ca+/calmodulin-dependent protein kinase. Plant Physiol 145:1619–1628

Chen C, Ane JM, Zhu H (2008) OsIPD3, an ortholog of the Medicago truncatula DMI3 interacting protein IPD3, is required for mycorrhizal symbiosis in rice. New Phytol 180:311–315

Craft D, Bangert M, Long T, Papp D, Unkelbach J (2014) Shared data for intensity modulated radiation therapy (IMRT) optimization research: the CORT dataset. GigaScience 3:37. doi:10.1186/2047-217X-3-37

Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, Liao X, Djari A, Rodriguez SC, Grohs C, Esquerré

D, Bouchez O, Rossignol MN, Klopp C, Rocha D, Fritz S, Eggen A, Bowman PJ, Coote D, Chamberlain AJ, Anderson C, Van Tassell CP, Hulsegge I, Goddard ME, Guldbrandtsen B, Lund MS, Veerkamp RF, Boichard DA, Fries R, Hayes BJ (2014) Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet 46:858–865. doi:10.1038/ng.3034

Dutkowski J, Kramer M, Surma MA, Balakrishnan R, Cherry JM, Krogan NJ, Ideker T (2013) A gene ontology inferred from molecular networks. Nat Biotechnol 31:38–45

Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich M, Mason BA, Goddard ME (2012) Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high density single nucleotide polymorphism panels. J Dairy Sci 95:4114–4129

Feuillet C, Stein N, Rossini L, Praud S, Mayer K, Schulman A, Eversole K, Appels R (2012) Integrating cereal genomics to support innovation in the Triticeae. Funct Integr Genomics 12:573–583. doi:10.1007/s10142-012-0300-5

Goddard ME (2014) Genetic architecture and evolution of quantitative traits. Proc Assoc Advmt Anim Breed Genet 20:122–125

Grant PR, Grant BR (2010) Conspecific versus heterospecific gene exchange between populations of Darwin's finches. Philosophical transactions of the Royal Society of London. Series B, Biol Sci 365:1065–1076

Gutjahr C, Banba M, Croset V, An K, Miyao A, An G, Hirochika H, Imaizumi-Anraku H, Paszkowski U (2008) Arbuscular mycorrhiza-specific signaling in rice transcends the common symbiosis signaling pathway. Plant Cell 20:2989–3005. doi:10.1105/tpc.108.062414

Ha M, Lu J, Tian L, Ramachandran V, Kasschau KD, Chapman EJ, Carrington JC, Chen X, Wang XJ, Chen ZJ (2009) Small RNAs serve as a genetic buffer against genomic shock in Arabidopsis interspecific hybrids and allopolyploids. Proc Natl Acad Sci U S A 106:17835–17840. doi:10.1073/pnas.0907003106

Haile-Mariam M, Nieuwhof GJ, Beard KT, Konstatinov KV, Hayes BJ (2013) Comparison of heritabilities of dairy traits in Australian Holstein-Friesian cattle from genomic and pedigree data and implications for genomic evaluations. J Anim Breed Genet 130:20–31. doi:10.1111/j.1439-0388.2013.01001.x

International Wheat Genome Sequencing Consortium (IWGSC) (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 345. doi:10.1126/science.1251788

Hofree M, Shen JP, Carter H, Gross A, Ideker T (2013) Network-based stratification of tumor mutations. Nat Methods 10:1108–1115. doi:10.1038/nmeth.2651

Kistner C, Parniske M (2002) Evolution of signal transduction in intracellular symbiosis. Trends Plant Sci 7:511–518

Kistner C, Winzer T, Pitzschke A, Mulder L, Sato S, Kaneko T, Tabata S, Sandal N, Stougaard J, Webb KJ, Szczyglowski K, Parniske M (2005) Seven Lotus japonicus genes required for transcriptional reprogramming of the root during fungal and bacterial symbiosis. Plant Cell 17:2217–2229

Kramer M, Dutkowski J, Yu M, Bafna V, Ideker T (2014) Inferring gene ontologies from pairwise similarity data. Bioinformatics 30:i34–42. doi:10.1093/bioinformatics/btu282

Krüttli A, Bouwman A, Akgül G, Della Casa P, Rühli F, Warinner C (2014) Ancient DNA analysis reveals high frequency of European lactase persistence allele (T-13910) in medieval Central Europe. PLoS One 9(1):e86251

Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW (2015). Insights from twenty years of bacterial genome sequencing. Funct Integr Gen 5. doi:10.1007/s10142-015-0433-4

Lee VA, Lorenz K (1978) The nutritional and physiological impact of milk in human nutrition. CRC Crit Rev Food Sci Nutr 11:41–116

Li B, Zhang G, Willerslev E, Wang J, Wang J (2011) Genomic data from the polar bear (Ursus maritimus). Available from: doi:10.5524/100008. GigaScience

Luo MC, Gu YQ, You FM, Deal KR, Ma Y, Hu Y, Huo N, Wang Y, Wang J, Chen S, Jorgensen CM, Zhang Y, McGuire PE, Pasternak S, Stein JC, Ware D, Kramer M, McCombie WR, Kianian SF, Martis MM, Mayer KF, Sehgal SK, Li W, Gill BS, Bevan MW, Simková H, Dolezel J, Weining S, Lazo GR, Anderson OD, Dvorak J (2013) A 4-gigabase physical map unlocks the structure and evolution of the complex genome of Aegilops tauschii, the wheat D-genome progenitor. Proc Natl Acad Sci U S A 110:7940–7945. doi:10.1073/pnas.1219082110

Mitra K, Carvunis AR, Ramesh SK, Ideker T (2013) Integrative approaches for finding modular structure in biological networks. Nat Rev Genet 14:719–732. doi:10.1038/nrg3552, Review

Monti S, Tamayo P, Mesirov J, Golub T (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn 52:91–118

Novarino G, Fenstermaker AG, Zaki MS, Hofree M, Silhavy JL, Heiberg AD, Abdellateef M, Rosti B, Scott E, Mansour L, Masri A, Kayserili H, Al-Aama JY, Abdel-Salam GM, Karminejad A, Kara M, Kara B, Bozorgmehri B, Ben-Omran T, Mojahedi F, Mahmoud IG, Bouslam N, Bouhouche A, Benomar A, Hanein S, Raymond L, Forlani S, Mascaro M, Selim L, Shehata N, Al-Allawi N, Bindu PS, Azam M, Gunel M, Caglayan A, Bilguvar K, Tolun A, Issa MY, Schroth J, Spencer EG, Rosti RO, Akizu N, Vaux KK, Johansen A, Koh AA, Megahed H, Durr A, Brice A, Stevanin G, Gabriel SB, Ideker T, Gleeson JG (2014) Exome sequencing links corticospinal motor neuron disease to common neurodegenerative disorders. Science 343:506–511. doi:10.1126/science.1247363

O'Bleness M, Searles VB, Dickens CM, Astling D, Albracht D, Mak AC, Lai YY, Lin C, Chu C, Graves T, Kwok PY, Wilson RK, Sikela JM (2014) Finished sequence and assembly of the DUF1220-rich 1q21 region using a haploid human genome. BMC Genomics 15:387. doi:10.1186/1471-2164-15-387

Oldroyd GE, Murray JD, Poole PS, Downie JA (2011) The rules of engagement in the legume-rhizobial symbiosis. Annu Rev Genet 45:119–144. doi:10.1146/annurev-genet-110410-132549

Qin JJ et al (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464:59–70. doi:10.1038/nature08821

Ramachandran V, Chen X (2008) Degradation of microRNAs by a family of exoribonucleases in Arabidopsis. Science 321:1490–1492. doi:10.1126/science.1163728

Reinhardt D, Kuhlemeier C (2002) Plant architecture. EMBO Rep 3:846–851

Ren G, Chen X, Yu B (2014) Small RNAs meet their targets: when methylation defends miRNAs from uridylation. RNA Biol 11(9):1099–1104. doi:10.4161/rna.36243

Rival P, de Billy F, Bono J-J, Gough C, Rosenberg C, Bensmihen S (2012) Epidermal and cortical roles of *NFP* and *DMI3* in coordinating early steps of nodulation in *Medicago truncatula*. Development 139:3383–3391. doi:10.1242/dev.081620

Rothammer S, Seichter D, Forster M, Medugorac I (2013) A genome-wide scan for signatures of differential selection in ten cattle breeds. BMC Genomics 14:908–925

Saatchi M, Beever JE, Decker JE, Faulkner DB, Freetly HC, Hansen SL, Yampara-Iquise H, Johnson KA, Kachman SD, Kerley MS, Kim J, Loy DD, Marques E, Neibergs HL, Pollak EJ, Schnabel RD, Seabury CM, Shike DW, Snelling WM, Spangler ML, Weaber RL, Garrick DJ, Taylor JF (2014) QTLs associated with dry matter intake, metabolic mid-test weight, growth and feed efficiency have little overlap across 4 beef cattle studies. BMC Genomics 15:1004. doi:10.1186/1471-2164-15-1004

Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, Pico AR, Bader GD, Ideker T (2012) A travel guide to Cytoscape plugins. Nat Methods 9(11):1069–1076. doi:10.1038/nmeth.2212

Shen B, Goodman HM (2004) Uridine addition after microRNA-directed cleavage. Science 306:997

Teeter KC, Payseur BA, Harris LW et al (2008) Genome-wide patterns of gene flow across a house mouse hybrid zone. Genome Res 18:67–76

van Doorn NLN, Wilson J, Hollund H, Soressi M (2012) Collins MJM (2012). Rapid Commun Mass Spectrom 26:2319

Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R (2010) Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol 6:e1000641

Wang J et al (2006) Genome wide non-additive gene regulation in Arabidopsis allotetraploids. Genetics 172:507–517

Warinner C, Hendy J, Speller C, Cappellini E, Fischer R, Trachsel C, Arneborg J, Lynnerup N, Craig OE, Swallow DM, Fotakis A, Christensen RJ, Olsen J, Liebert A, Montalva N, Fiddyment S, Mackie M, Canci A, Bouwman A, Rühli F, Gilbert MTP, Collins MJ (2014a) Direct evidence of milk consumption from ancient human dental calculus. Scientific Reports 4:7104. doi:10.1038/srep07104

Warinner C, Speller C, Collins MJ (2014b) A new era in paleomicrobiology: prospects for ancient dental calculus as a long-term record of the human oral microbiome. Philos Trans R Soc B 370:20130376. doi:10.1098/rstb.2013.0376

Warinner C et al (2014c) Ancient human microbiomes. J Hum Evol. doi:10.1016/j.jhevol.2014.10.016

Warinner C, Rodrigues JFM, Vyas R, Trachsel C, Shved N, Grossmann J, Radini A, Hancock Y, Tito RY, Fiddyment S, Speller C, Hendy J, Charlton S, Luder HU, Salazar-García DC, Eppler E, Seiler R, Hansen L, Samaniego Castruita JA, Barkow-Oesterreicher S, Teoh KY, Kelstrup C, Olsen JV, Nanni P, Kawai T, Willerslev E, von Mering C, Lewis CM Jr, Collins MJ, Gilbert MTP, Rühli F, Cappellini E (2014d) Pathogens and host immunity in the ancient human oral cavity. Nat Genet 46(4):336–344. doi:10.1038/ng.2906

Xie F, Murray JD, Kim J, Heckmann AB, Edwards A, Oldroyd GE, Downie JA (2012) Legume pectate lyase required for root infection by rhizobia. Proce Nat Acad Sci USA 109(2):633–638

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders A, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW et al (2010) Nature Genet 42:565

Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, de Andrade M, Feenstra B, Feingold E, Hayes MG et al (2011) Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet 43:519–525

Yang SY, Grønlund M, Jakobsen I, Grotemeyer MS, Rentsch D, Miyao A, Hirochika H, Kumar CS, Sundaresan V, Salamin N, Catausan S, Mattes N, Heuer S, Paszkowski U (2012) Nonredundant regulation of rice arbuscular mycorrhizal symbiosis by two members of the phosphate transporter1 gene family. Plant Cell 24:4236–4251. doi:10.1105/tpc.112.104901

Yu B, Yang Z, Li J, Minakhina S, Yang M, Padgett RW, Steward R, Chen X (2005) Methylation as a crucial step in plant microRNA biogenesis. Science 307:932–935, PubMed: 15705854

Yumul RE, Kim YJ, Liu X, Wang R, Ding J et al (2013) POWERDRESS and diversified expression of the MIR172 gene family bolster the floral stem cell network. PLoS Genet 9(1):e1003218. doi:10.1371/journal.pgen.1003218