Funct Integr Genomics (2009) 9:1–6
DOI 10.1007/s10142-009-0112-4

EDITORIAL

Diversity of genome research at the 2009 Plant and Animal Genome Conference

R. Appels

Keywords Genome research · Technology · Integration · Application

The annual Plant and Animal Genome (PAG) Conference held in San Diego, 10th–14th January 2009, was a large meeting (2,300 participants) covering the diversity of science underpinning genome-level analyses in both the commercial and fundamental areas. The structure of the meeting was participant-driven through 105 workshops spread over 5 days. The conference was, as always, extremely well organized and included plenary lectures to define major issues and provide insights into the future directions of biotechnology in its broadest definition. This editorial is largely focused on the plenary lectures and is clearly a personal interpretation of the meeting and lectures, validated where possible using recent publications, and is not intended to imply approval from the PAG or presenters at the conference.

The new technologies

The new generation of sequencing technologies formed the basis of many presentations in the workshops. Outputs focused on developing high-throughput assays of single-nucleotide polymorphisms in microarray formats in order to identify genetic markers for assaying variation within domesticated plants and animals for breeding as well as in fundamental studies in model organisms such as *Drosophila* and *Arabidopsis*. The range of microbial, plant, insect, and animal life-forms that are now analyzed using the new DNA and RNA sequencing technologies was very extensive. In medicine, the ambition to monitor molecular-level information in individuals was driving advances in molecular biotechnology, chemistry, and microengineering.

G Church highlighted the four-log improvement in DNA technology over the past 4 years (*Polonator.org*; Shendure et al. 2005) with respect to reductions in costs, increased speed, and improvements in interpretability. He discussed targeted sequencing of coding variants (~1% of the genome) plus analysis of regulatory variants through quantifying RNA levels using sequence tag counting. In the area of personalized medicine, Church described the monitoring of environmental components via the microbiome and its interaction with human individuals via the blood-immunoglobulin (VDJ)-ome, through haplotype studies and allele-specific expression. The latter method was argued to provide new data for establishing causative links. To date, a personal genome showed approximately 10,000 DNA variations which affect protein structure and three million which do not and Church argued that association studies of common DNA variants with diseases often yielded weak predictive power and few causative alleles. The efforts in genome-wide sequencing and the aggregation of alleles in specific gene/pathway functions were predicted to significantly impact preventative medicine and more accurate diagnostics. The large personalized medicine project (*PersonalGenomes.org*) currently being developed was argued to be a unique open-access effort to integrate personal genomes with comprehensive sets of medical and non-medical traits and environmental measures.

L Hood outlined the advances in click chemistry as a modular approach to bioconjugation reactions in both proteomics and DNA research. The copper-(I)-catalyzed 1,2,3-triazole formation from azides and terminal acetylenes was the linking reaction of choice due to its high

R. Appels (✉)
Center for Comparative Genomics, Murdoch University,
Perth, Western Australia 6150, Australia
e-mail: rapp1495@bigpond.net.au

degree of dependability and specificity (reviewed in Kolb and Sharpless 2003). The azides are formed on cysteine and lysine residues in proteins as precursors to the triazole formation. The triazole products also associate with biological targets, through hydrogen bonding and dipole interactions. It was evident that the concepts of bead/array-based chemistry that have underpinned the revolution in DNA and RNA sequencing have wider applications as the click chemistries provide the basis for the construction of extensive protein matrices. The protein matrix technology was discussed in the context of identifying specific blood proteins in particular, in order to provide biomarkers that can be utilized for disease diagnosis or for assessing the impact of drugs treatments in individuals. Hood indicated that advances in a range of technologies and resources such as next-generation DNA sequencing, targeted mass spectrometry, microfluidic chips, and DNA-encoded antibody libraries (Bailey et al. 2007) would enable the implementation of a personalized medicine that was "predictive, preventive, personalized and participatory" (P4 medicine, Hood 2008). Engagement in personalized medicine programs was seen to be a new and challenging phase for biotechnology and matching progress in education was argued by Hood to be crucial as individuals needed to understand their own records and interventions aimed at reducing the risk factors affecting their health.

The computational challenges generated by the new-generation sequencing technologies were detailed by E Birney. Birney introduced Ensembl (www.ensembl.org) in its role of contributing to the complex process of pattern matching of protein and DNA sequences (software "pipelines" written in Perl). Open source and public accessibility underpinned the philosophy of Ensembl. The expanding remit of Ensembl underscored the speed of the advances occurring in the genomics area: *mammals* (*primates*: bush baby, chimp, human, macaque, mouse lemur, orangutan, tarsier; *rodents*: guinea pig, kangaroo rat, mouse, pika, rabbit, rat, ground squirrel, tree shrew; *Laurasiatheria*: alpaca, cat, cow, dog, dolphin, hedgehog, horse, megabat, microbat, shrew, pig; *Afrotheria*: elephant, hyrax, tenrec; *Xenarthra*: armadillo); *marsupials and monotremes* (opossum, platypus); *birds* (chicken); fish (fugu, green-spotted puffer fish, zebra fish, medaka, stickleback, sea lamprey); *reptiles and amphibians* (*Xenopus*, anole lizard; ancient relatives: *Ciona intestinalis*, *Ciona savignyi*); *invertebrates* (*insects*: mosquito (two species), fruit fly; *worm*: *Caenorhabditis elegans*); *yeast* (baker's yeast). For the assembly of the short reads generated by next generation sequencing technologies, Birney described the software Velvet (Zerbino and Birney 2008) which used de Bruijn graphs. He explained that the fundamental data structure in the de Bruijn graph was based on *k*mers, rather than the reads from sequencing outputs, and dealt more effectively with the high redundancy in datasets and with the defined start and end points in the paired reads methodology. Errors in assembly were apparently more easily avoided and searches for overlaps were simplified. Birney also addressed a recurrent theme in the conference, namely, the annotation of the genome sequence (Encode Consortium 2007). In particular, he emphasized the need for (1) deep RNA sequencing in order to accurately define the regions of the genome that are transcribed; (2) sequence-based comparative genomics to define features such as duplications; (3) the location of DNA methylation and DNase hypersensitive sites (a feature of the chromatin assembly); (4) distribution of histone modifications as they relate to the genome sequence (a feature of the chromatin assembly). With regards to computing technologies *per se*, Birney emphasized that the available memory (rather than storage space) is fast becoming a bottleneck in the assembly of large genomes using the next generation sequencing technologies.

The biological impacts and the integration challenge

The report on the human physiome project in PAG XV (Hunter 2007; Hunter and Borg 2003) noted the complexity of biological systems and that the vast amount of information now available at the level of genes, proteins, cells, tissues, and organs required the development of models to define the relationship between structure and function at all levels of biological organization. Figure 1 shows a visual translation of this concept to cereals and is equally applicable to any biological system.

In the period of time since Hunter and Borg (2003), there has been a revolution in the pace at which DNA and RNA sequences can be generated. In his talk, L Hood argued that biology, as an informational science (systems biology), was now the driver for the integration of innovation in computational, chemical, and instrumentation disciplines. He used his group's analysis of the prion disease in mice to illustrate the combination of large-scale screening processes to identify all the prion-related genes (Hood 2008), building on earlier detailed genome-level analyses (Lee et al. 1998) and an understanding of the biology (Moore et al. 2001; Hood 2008). Monitoring the environment through characterizing oral, gut, and environmental bacteria was feasible using the new sequencing technologies and G Church reported that the high levels of antibiotic resistance that are now being found may eventually necessitate alternative approaches to combating unwanted life-forms in certain situations. Allele-replacement strategies to remove the dependence on, for example, a particular transfer RNA for protein synthesis (an invading life-form would still be dependent on this basic entity) was discussed in terms of utilizing defects in the methyl-directed mismatch repair system operating at the replication fork during DNA

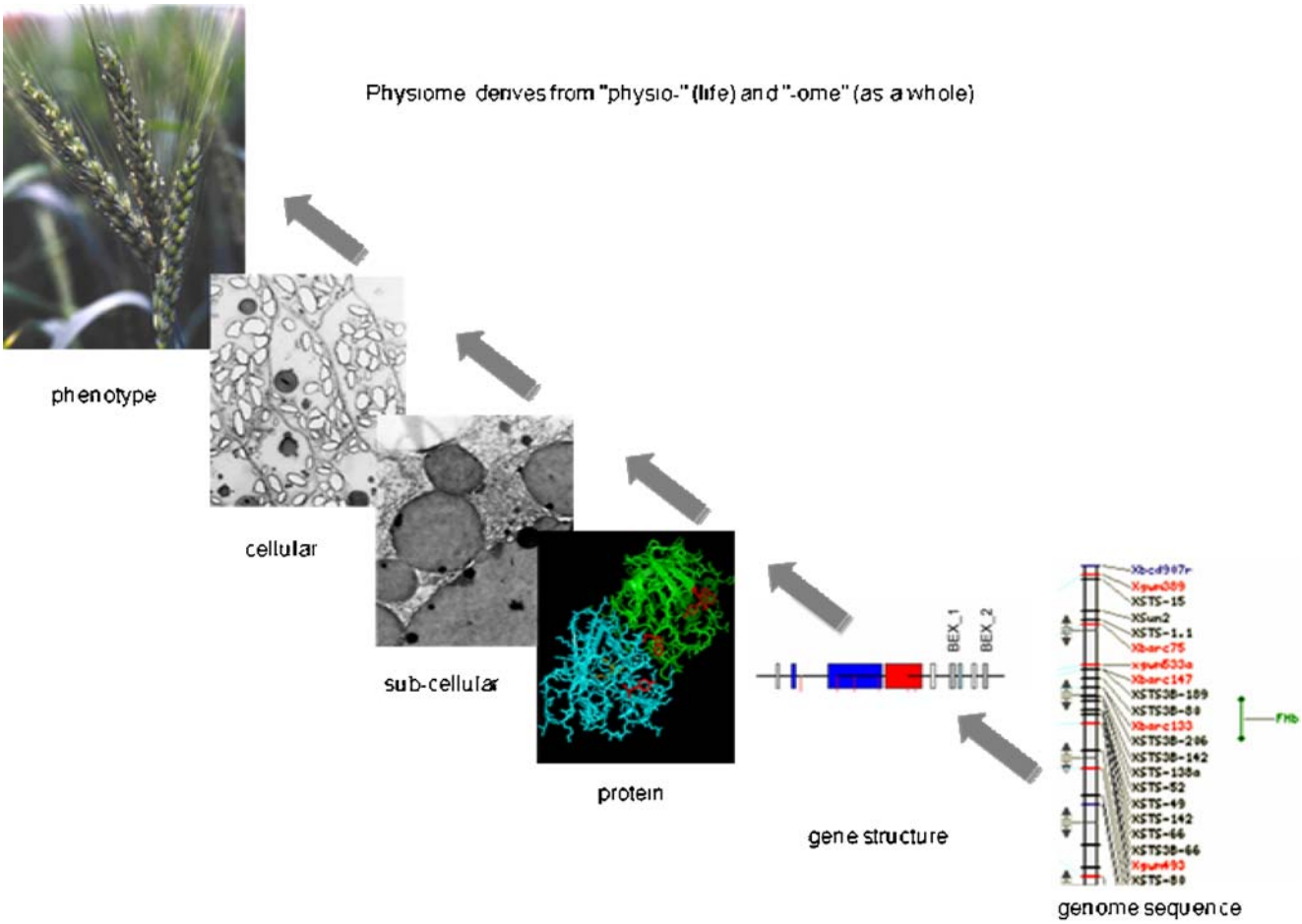

Fig. 1 The flow of information for the cereal physiome has been modeled after the human physiome report (R. Appels unpublished)

synthesis (Costantino and Court 2003) on a large scale (Forster and Church 2007). Novel sources of resistance of proteins to enzymes, parasite, and predators were also discussed by Church in terms of the requirements for producing proteins using D-amino acids instead of the usual L-stereoisomers of amino acids that occur in the proteins of life-forms. The production of so-called mirror proteins necessitated the re-engineering of many fundamental aspects of the protein synthesis machinery. The overall concept of modifying life-forms at a very basic level was also envisaged as a way forward to increase photosynthesis efficiencies and to develop biofuels from algae and grasses.

The correct assignment of gene sequences within the genome is fundamental to integration of the DNA-encoded information and, as S Briggs described it, is also the basis for new discoveries. One key cross-reference in the correct annotation of the genome was that of the proteome back to the hypothesized gene sequences in order to achieve the required validation. Although this has been carried out in a targeted way (Finnie et al. 2008; Kiel et al. 2009), Briggs discussed the results of large-scale sampling of peptides and the process of relating these to gene sequences in the genome. The technology base for the large-scale analysis of the *Arabidopsis* proteo-genome used tandem mass spectroscopy as well as a diversity of tissue samples (Castellana et al. 2008). The identification of peptide sequences was typically based on relating the exact molecular weight of a particular peptide to a database of known peptides to determine the sequence. Post-translational modifications that change the mass and pattern of peptides were also taken into account by capturing subgroups of peptides carrying either phosphate or acetyl groups. The amino acid sequences of 144,079 distinct peptides were determined and the majority of the peptides (126,055) resided in existing gene models (12,769 confirmed proteins), comprising 40% of annotated genes. The remaining peptides (18,024) did not correspond to annotated genes and formed the basis for correcting the annotation of many genes in the genome. Within this group, 778 new protein-coding genes were discovered (thylakoid lumen protein gene was specifically mentioned). Three peptides were located within a genome region classified as a transposable element and the gene corresponding to these peptides had high

similarity to the ubiquitin-like protease (Ulp1) family in *Arabidopsis*. These findings have significance for large genomes such as wheat because of their much higher content of retrotransposable elements. For the human proteo-genome (Tanner et al. 2007), 39,000 exons were validated and the translation of 224 hypothetical proteins was confirmed. Novel exons were identified for 16 genes and over 40 alternative splicing events were identified. Briggs noted that among the variables defining a gene (translation start, splice boundaries, frameshifts, alternative splicing), frameshifts were a key issue.

S Briggs noted that comparative proteo-genomic studies provided important evidence for cross-referencing the proteome to genome annotation (Gupta et al. 2008). In *Drosophila*, A Clark reviewed the results from the "12 *Drosophila* genomes project," to provide an insight into the contributions that a large-scale study of this type can make to the annotation of a genome (Stark et al. 2007). Using the color coding of triplets to differentiate changes that preserve the protein sequences from others that are typical of non-coding regions, a good visualization was generated for a detailed analysis that identified 149 genes with apparent stop-codon read through and 123 novel polycistronic transcripts as well as many refinements to the existing annotation of the *Drosophila* genome. The extensive comparative genomics approach also allowed the transcriptome to be defined with respect to small RNAs and their targets and adenosine-to-inosine RNA editing. Clark argued that the phylogenetic-based analysis provided the basis for determining rates of substitution, gene family expansion, and genome rearrangements. Clark went on to examine how natural selection has affected patterns of gene family evolution and sequence divergence among different components of the innate immune system, with a particular focus on drosomycin, relish, and cecropin. As was the case for the genes identified in Sackton et al. (2008), these immune-system genes evolved under positive Darwinian selection. Positively selected sites within recognition proteins clustered in domains involved in the recognition of microorganisms, suggesting that molecular interactions between hosts and pathogens may have driven adaptive evolution in the *Drosophila* immune system. It was evident that *Drosophila* has also been a primary model organism for population genetic studies, especially in relation to demonstrating the utility of linkage disequilibrium. Clark made the case for *Drosophila* providing a model for understanding differences in risk associated with complex traits between individuals of a population, after defined genes have been identified as being linked to the respective trait through the analysis of allelic variants.

The increasingly powerful technologies evident at the PAG have also been applied to polyploids in plant groups, and the resulting discoveries have started to define novel genomic interactions. The developments in *Gossypium* were discussed by J Wendel (Iowa State University). It was evident that the merger of diverse genomes in *Gossypium* to form polyploids, some 1–2 MYA, generated a spectrum of responses including disruption and reconciliation of ancestral gene expression patterns. Using several microarray platforms (Udall et al. 2007), global transcriptional changes in synthetic and natural *Gossypium* allopolyploids and reconstructed F1 and polyploid hybrids were determined for different tissues and genetic backgrounds. Allopolyploid formation was found to have induced massive alterations in gene expression and complex transcriptomic responses, including genome-wide genomic dominance and novel (transgressive) expression patterns. Using a novel microarray that simultaneously distinguished transcript levels for each homoeolog, Wendel demonstrated that allopolyploidization entailed significant homoeolog expression modulation that was temporally partitioned into alterations arising immediately as a consequence of genomic merger and secondarily as a result of long-term evolutionary transformations in duplicate gene expression. The complex gene networks controlling, for example, lint fiber development (Rong et al. 2007) have meant that expression in some tissues may be biased such that there is an overall unequal contribution of two genomes to the transcriptome. Homoeolog expression ratios changed during fiber development, showing that duplicate gene expression modulation characterized the development of a single cell. The functional consequences of gene duplication in cotton and the possibility of novel gene recruitment following genome doubling were defined as areas for future research.

Advances in polyploid wheat provided another example of the significant progress being made in organisms previously considered "too hard" (Paux et al. 2008). Similar to cotton, wheat showed homoeologous gene silencing (Mochida et al. 2003; Zhang et al. 2008; Bottley and Koebner 2008). Bottley and Koebner (2008) identified homoeolog non-expression for 15 single-copy genes across a panel of 16 wheat varieties and found the expression profiles of eight genes indicating that only two varieties shared the same pattern of silencing. Since epigenetic variation exerted a significant effect on phenotype, it was argued that the ubiquity and variability in homoeologous silencing observed in wheat was likely to be significant in generating phenotypic variation. In the area of identifying genes underpinning resistance to diseases, genome-level sequencing in wheat provided the basis for delineating the molecular nature of the Tsn1 (J Faris), Lr34 (E Lagudah), Yr36 (J Dubcovsky), and Sr2 (W Spielmeyer), genes that provide sensitivity to *Stagonospora* toxin and resistance to leaf, stripe, and stem rusts, respectively. These genes represent important tools in breeding in order to maintain yield in many environments where these pathogens can

cause major losses. Studies on the Tsn1, Lr34, and Yr36 genes, in particular, have demonstrated the different molecular mechanisms for resistance to fungal diseases. The Tsn1 gene encoded an NBS-LRR protein, consistent with gene structures found in other resistance genes (e.g., Rpg5 in barley, Brueggeman et al. 2008). The Lr34 gene was found to be in the ABC transporter class of gene, postulated to be involved in secreting a currently undefined molecule to inhibit the growth of fungal hypha penetrating the leaf tissue. In contrast, the Yr36 encoded a protein with a novel architecture resulting from domain reshuffling involving a kinase and a putative lipid binding domain. The occurrence of a fusion of these two particular domains was, to date, only found in the Triticeae species. The occurrence of a fusion of domains of this type was analogous to the fusion product that was formed as a result of the translocation between chromosome-9 and chromosome-22, the Philadelphia chromosome, in humans. The translocation fused the break point cluster region sequences of chromosome-22 with the c-ABL tyrosine kinase of chromosome-9, replacing exon 1, to generate a chimeric gene product with a tyrosine kinase activity several fold higher than normal and which correlates with the disease phenotype (chronic myelodysplastic hematopoietic stem cell disorder, reviewed in Goodsell 2006).

The translation to wider challenges

A unique insight into the transfer of many aspects of the broad area of biotechnology into a commercial production line was provided by D Aviezer, from ProCellEx. The projects discussed in detail related to the plant-cell-based production of proteins/enzymes with therapeutic applications and were thus particularly interesting in that they spanned the plant–animal divide. The adaption of gene constructs for high level and stable expression in plant cell culture was demonstrated using carrot cells. Cell cultures were grown in flexible, sterile, polyethylene bioreactors in chemically defined growth medium. The enzyme used in the treatment of Gaucher disease, glucocerebrosidase (also known as acid β-glucosidase), was obtained on a large-scale using the culture system and was found to have a very similar, although not identical, pattern of post-translation modification by mannose glycans relative to that produced in mammalian cells. The product was competitive with the protein expressed in Chinese hamster (CH) cell lines in terms of the 3D structure and performance in trials, even though the glycosylation pattern of the final product was not identical with that from CH cells (Shaaltiel et al. 2007). Several other proteins, including acetylcholine esterase, were indicated to be in production.

The impacts of biotechnology concepts and technologies in non-commercial agricultural areas were considered by R Horsch. The big challenges in agricultural sciences were clearly articulated and included water conservation, soil fertility, resistance of domesticated plants and animals to pests and diseases, the production of quality grain from resilient, productive, and adapted varieties, best management practices, and managing climate change. Solutions involved on the one hand managing resources so that scale of effort was proportionate to need and opportunity. Horsch made the point that communities in need of innovation and improved management of resources were usually risk averse which resulted in a reluctance to invest in change and hence progress was often limited. Conventional strategies such as the use of improved hybrid seeds and fertilizers, educating farmers, and increasing their access to markets remained in the forefront for helping the poorest of the world's poor. On the other hand, there was also a need to capture biotechnological breakthroughs that can radically change the phenotypes of domesticated plants and animals for maximum performance in specified environments. The shortage of phosphorus (Cordell 2008) was an example of where engineering solutions were needed to provide breakthroughs in extracting phosphorus from a range of biological waste products while breeding/biotechnology solutions were needed to select new genotypes that use phosphorus more efficiently. Speed in delivering new genotypes as grain for planting was crucial and marker-assisted selection applications based on more informative diagnostic probes to identify favorable combinations of alleles and genome segments in new germplasm remained a high priority. Wide crossing and GM programs were currently the main sources of new genetic variation. In the case of an emerging threat to wheat in the form of the stem rust Ug99, new sources of resistant phenotypes were urgently needed. One option being considered was rebuilding the "Sr2 complex" with other unknown additive genes of similar nature to achieve long-term resistance (Singh et al. 2006) and this process would be accelerated by the new insights into the molecular basis for disease resistance in wheat. Increased efficiencies in photosynthesis and water use, as well as the incorporation of nutritional attributes, may need more fundamental changes to the genomes of mainstream cereals in order to provide grain for future generations.

The area of personalized medicine represented a major translation of a wide range of knowledge, technologies, and microengineering into everyday use and both L Hood and G Church provided unique insights into the new concepts and drivers in this area. The great many researchers not working in the medical area, or with model organisms, were offered a glimpse into the future with respect to concepts and technologies in their own particular area that will undoubtedly follow in the footsteps of investment in biotechnology and systems biology in other life-forms.

References

Bailey RC, Kwong GA, Radu GG, Witte ON, Heath JR (2007) DNA-encoded antibody libraries: a unified platform for multiplexed cell sorting and detection of genes and proteins. J Am Chem Soc 129:1959–1967

Bottley A, Koebner RM (2008) Variation for homoeologous gene silencing in hexaploid wheat. Plant J 56:297–302

Brueggeman R, Druka A, Nirmala J, Cavileer T, Drader T, Rostoks N, Mirlohi A, Bennypaul H, Gill U, Kudrna D, Whitelaw C, Kilian A, Han F, Sun Y, Gill K, Steffenson B, Kleinhofs A (2008) The stem rust resistance gene Rpg5 encodes a protein with nucleotide-binding-site, leucine-rich, and protein kinase domains. Proc Natl Acad Sci 105:14970–14975

Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V, Steven P, Briggs SP (2008) Discovery and revision of *Arabidopsis* genes by proteogenomics. Proc Natl Acad Sci 105:21034–21038

Cordell D (2008) The story of phosphorus: missing global governance of a critical resource. SENSE Earth Systems Governance, Amsterdam, 24th–31st August

Costantino N, Court DL (2003) Enhanced levels of λ red-mediated recombinants in mismatch repair mutants. Proc Natl Acad Sci 100:15748–15753

ENCODE Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816

Finnie C, Bagge M, Steenholdt T, Østergaard O, Bak-Jensen KS, Backes G, Jensen A, Giese H, Larsen J, Roepstorff P, Svensson B (2008) Integration of the barley genetic and seed proteome maps for chromosome 1H, 2H, 3H, 5H and 7H. Funct Integr Genomics. doi:10.1007/s10142-008-0101-z

Forster AC, Church GM (2007) Synthetic biology projects in vitro. Genome Res 17:1–6

Goodsell DS (2006) The molecular perspective: c-Abl tyrosine kinase. Stem Cells 24:209–210

Gupta N, Benhamida J, Bhargava V, Goodman D, Kain E, Kerman I, Nguyen N, Ollikainen N, Rodriguez J, Wang J, Lipton MS, Romine M, Bafna V, Smith RD, Pevzner PA (2008). Comparative proteogenomics: Combining mass spectrometry and comparative genomics to analyze multiple genomes. Genome Res 18:1133–1142

Hood L (2008) A personal journey of discovery: developing technology and changing biology. Ann Rev Anal Chem 1:1–43

Hunter PJ (2007) The IUPS physiome project: a framework for multi-scale computational physiology. In: Plant & Animal Genome (PAG) XV Conference

Hunter PJ, Borg TK (2003) Integration from proteins to organs: the physiome project. Nat Rev Mol Cell Biol 4:237–243

Kiel JAKW, van den Berg MA, Fusetti F, Poolman B, Bovenberg RAL, Veenhuis M, van der Klei IJ (2009) Matching the proteome to the genome: the microbody of penicillin-producing *Penicillium chrysogenum* cells. Funct Integr Genomics 9: doi:10.1007/S10142-009-0110-6

Kolb HC, Sharpless KB (2003) The growing impact of click chemistry on drug discovery. Drug Discov Today 8:1128–1137

Lee IY, Westaway D, Smit AFA, Wang K, Seto J, Chen L, Acharya C, Ankener M, Baskin D, Cooper C, Yao H, Prusiner SB, Hood L (1998) Complete genomic sequence and analysis of the prion protein gene region from three mammalian species. Genome Res 8:1022–1037

Mochida K, Yamazaki Y, Ogihara Y (2003) Discrimination of homoeologous gene expression in hexaploid wheat by SNP analysis of contigs grouped from a large number of expressed sequence tags. Mol Gen Genomics 270:371–377

Moore RC, Xiang F, Monaghan J, Han D, Zhang Z, Edstrom L, Anvret M, Prusiner SB (2001) Huntington disease phenocopy is a familial prion disease. Am J Hum Genet 69:1385–1388

Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, Leroy P, Korol A, Michalak M, Kianian S, Spielmeyer W, Lagudah E, Somers D, Kilian A, Alaux M, Vautrin S, Bergès H, Eversole K, Appels R, Safar J, Simkova H, Dolezel J, Bernard M, Feuillet C (2008) A physical map of the 1-gigabase bread wheat chromosome 3B. Science 322:101–104

Rong J, Feltus FA, Waghmare VN, Pierce GJ, Chee PW, Draye X, Saranga Y, Wright RJ, Wilkins TA, May OL, Smith CW, Gannaway JR, Wendel J, Paterson A (2007) Meta-analysis of polyploid cotton QTL shows unequal contributions of subgenomes to a complex network of genes and gene clusters implicated in lint fiber development. Genetics 176:2577–2588

Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG (2008) Dynamic evolution of the innate immune system in *Drosophila*. Nat Genet 39:1461–1468

Shaaltiel Y, Bartfeld D, Hashmueli S, Baum G, Brill-Almon E, Galili G, Dym O, Boldin-Adamsky SA, Silman I, Sussman JL, Futerman AH, Aviezer D (2007) Production of glucocerebrosidase with terminal mannose glycans for enzyme replacement therapy of Gaucher's disease using a plant cell system. Plant Biotechnol J 5:579–590

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Kang K, Mitra RD, Church GM (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732

Singh RP, Hodson DP, Jin Y, Huerta-Espino J, Kinyua MG, Wanyera R, Njau P, Ward R (2006) Current status, likely migration and strategies to mitigate the threat to wheat production from race Ug99 (TTKS) of stem rust pathogen. CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources 2006 1, No. 054: 1–13

Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, Ruby JG, Brennecke J, Harvard FlyBase curators, Berkeley Drosophila Genome Project, Hodges E, Hinrichs AS, Caspi A, Paten B, Park S-W, Han MV, Maeder ML, Polansky BJ, Robson BE, Aerts S, van Helden J, Hassan B, Gilbert DG, Eastmann DA, Rice M, Weir M, Hahn MW, Park Y, Dewey CN, Pachter L, Kent WL, Haussler D, Lai E, Bartel DP, Hannon GJ, Kaufman TC, Eisen MB, Clark AG, Smith D, Celkiner SE, Gelbart WM, Kellis M (2007) Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. Nature 450:219–232

Tanner S, Shen Z, Ng J, Florea L, Guigó R, Briggs SP, Vineet Bafna V (2007) Improving gene annotation using peptide mass spectrometry. Genome Res 17:231–239

Udall JA, Flagel LE, Cheung F, Woodward AW, Hovav R, Rapp RA, Swanson JM, Lee JJ, Gingle AR, Nettleton D, Town C, Chen ZJ, Wendel JF (2007) Spotted cotton oligonucleotide microarrays for gene expression analysis. BMC Genomics 8:81–89

Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829

Zhang J, Huang S, Fosu-Nyarko J, Dell B, McNeil M, Waters I, Moolhuijzen P, Conocono E, Appels R (2008) The genome structure of the 1-FEH genes in wheat (*Triticum aestivum* L): new markers to track stem carbohydrates and grain filling QTLs in breeding. Mol Breeding 22:339–351