NATIONAL INSTITUTES OF BIOSCIENCE JOURNAL

Genome Editing and the Future of Farming

Conference held September 6th, 2016 at The Roslin Institute, Edinburgh

Genome reference quality an essential resource in an age of genome editing

Wesley C. Warren[1] and David W. Burt[2]

[1]McDonnell Genome Institute, Washington University School of Medicine, St Louis, MO 63108; [2]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian, EH25 9RG, Edinburgh, UK.

Corresponding author:
wwarren@wustl.edu

OECD
BETTER POLICIES FOR BETTER LIVES

The opinions expressed and arguments employed in this publication are the sole responsibility of the authors and do not necessarily reflect those of the OECD or of the governments of its Member countries.

The Conference was sponsored by the OECD Co-operative Research Programme on Biological Resource Management for Sustainable Agricultural Systems, whose financial support made it possible for some of the invited speakers to participate in the Conference.

Genome reference quality an essential resource in an age of genome editing

Wesley C. Warren[1] and David W. Burt[2]

[1]McDonnell Genome Institute, Washington University School of Medicine, St Louis, MO 63108; [2]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian, EH25 9RG, Edinburgh, UK.

Corresponding author:
wwarren@wustl.edu

ABSTRACT

With the recent exploration of how we may improve livestock production and meet growing demand for animal protein products using genome editing technology, we argue that exemplary genome references will be required to ensure that the proposed edits are specific and carefully evaluated for any potentially harmful side effects. We explore in this short review the status of existing genome references for the major food producing animals (cattle, chicken, pigs, goat and sheep) and summarise best practice for creating future higher quality genome references. Each will serve as a central conduit in the study of genetic manipulation outcomes, and provide a computational workflow for how the edited genome could be evaluated for no other unexpected base changes in the rest of the genome.

CONFERENCE PAPER

A significant contribution to experimental model systems permeates the history of domestic animal studies (Megens and Groenen 2012). Many reproductive success stories in human fertility were first pioneered in cattle; the transgenic cow is used to produce proteins in their milk for human therapeutics, and a long history of collecting animal tissues from abattoirs for the purification of various biologicals continues today.

 http://dx.doi.org/10.2218/natlinstbiosci.1.2016.1745

The finding that injections of tumour filtrate into healthy chickens reproduced observed tumours initiated the field of viral oncology (Rubin 2011). These are but a few examples that highlight the many contributions that food-producing animals have made to advances in biomedical science. However, their most significant contribution to society is as a food source; and suffice to say that without a safe and efficient supply of food-producing animals, a significant percentage of our world population would be severely malnourished and possibly starve to death. The ability to feed the world is even more urgent today, with a world population predicted to reach 9.7 billion by 2050 (UN DESA Report; https://esa.un.org/unpd/wpp/). With the recent exploration of how we may improve livestock production and meet growing demand for animal protein products using genome editing technology, we argue that exemplary genome references will be required to ensure that the proposed edits are specific and carefully evaluated for any potentially harmful side effects. We explore in this short review the status of existing genome references for the major food producing animals (cattle, chicken, pigs, goat and sheep), summarise best practice for creating future higher quality genome references. Each will serve as a central conduit in the study of genetic manipulation outcomes, and provide a computational workflow for how the edited genome could be evaluated for no other unexpected base changes in the rest of the genome.

Today we are fortunate to have access to sequencing technology that can advance our ability to obtain near complete DNA sequences of each food-producing genome. At present, moderate-quality genome references are available for all food-producing species including cattle, chicken, sheep and pig that can serve as a computational starting point to ensure the traits we wish to protect, enhance or suppress are studied with a relatively small loss of information (Table 1). We label these references as 'moderate' quality since the most realistic measure of completeness is the number of contigs or "gap-free sequences" being equal to the expected total chromosome count. However, in each case, total contig numbers are far higher in these animals compared to the human genome. Advances in sequencing technology and physical mapping of

chromosomes, specifically those producing longer reads, have brought on the eager expectation that we will elevate each of these references to near human quality, hopefully, single scaffolds per chromosome with a small number of contigs per scaffold. The recently assembled goat genome provides validation of this expectation with 31 assembled scaffolds equivalent to the expected number of chromosomes (Derek M Bickhart 2016). Moreover, we are aware of recent assemblies of the chicken, pig and bovine genomes using this same path of long read technology that promises to offer the community high genome reference quality for future computational and genomic studies.

A variety of approaches can be used to address sequence connectivity deficits observed in these genome references (Table 1). However, to date, the best practice for a vertebrate is first to sequence the genome to a minimum of 60x sequence coverage of long reads (mean size ~14kb) and assemble all reads with the best-suited algorithm. Once high molecular weight DNA (>50kb fragment length) is obtained (a crucial first step to success), to our knowledge, all vertebrate genomes are being sequenced on the PacBio RSII instrument with Single Molecule Real-Time (SMRT) reagents. It is likely that the recently introduced PacBio Sequel instrument will supplant the RSII as soon as read length reaches RSII equivalency or close to it. Currently, the RSII instrument routinely provides an average read length of ~14 kb in our production labs with the longest read lengths often exceeding 50 kb. Individual PacBio read error rates (~85%) are resolved by high sequence coverage (>50-fold), which allows generation of highly accurate base consensus (>99.9%). Long-read sequence assemblers continue to evolve, but we have adopted the use of the DALIGNER (Myers 2014) as a first step toward read error correction and FALCON for read overlap and string graph layout, followed by QUIVER to generate consensus error-corrected sequence (Chin 2014). Despite the efficiency of error correction with QUIVER, we have found it necessary to clean up residual errors, mostly insertions and deletions, using aligned Illumina paired-end (125bp length) sequences and PILON (Walker *et al*. 2014). False protein-coding frameshifts are largely eliminated as a result of this final step.

Using one such assembly algorithm (Berlin *et al*. 2015), the goat (*Capra hircus*) genome reached unprecedented levels of sequence continuity (Table 1), thus demonstrating the clear advantages of recent technological advances in genome assembly.

Starting with the most contiguous assembly possible, the next step is to apply high-resolution mapping/phasing technology, such as chromatin sequence maps that will produce a proximity-guided assembly, thus creating chromosome-scale scaffolds that in theory should match total chromosome count. Fortunately, recent methodological advances have mostly overcome prior assembly connectivity bottlenecks by either adapting a chromosome conformation capture technique (Selvaraj *et al*. 2013) or utilizing restriction enzyme cuts of long DNA strands that are separated on nanochannel arrays (Hastie *et al*. 2013). By using a combination of these scaffolding methods the 3,074 assembled goat contigs were connected to a final count of 31 scaffolds, the known number of chromosomes for goat (Bickhart 2016). In the chicken, genome-wide study designs continue to be incomplete due to missing autosomes, in particular, the high GC-content microchromosomes (cite G3 paper). Utilization of long read assemblies and high-resolution maps will resolve most of these deficiencies.

In the final phase of genome assembly curation, its accuracy is typically judged by the following metrics: the appearance of homologous reference differences compared to called single base, small (<6bp) insertion or deletions, all from same source DNA sequences, and if available long mate pair sequences that display alignment discordance in order or orientation of scaffolds or contigs within scaffolds. For the latter, a conundrum is few automated tools can make genome-wide decisions on assembly order or orientation without manual review as these often involve repeats, segmental duplications or tandem arranged gene families.

Over a decade ago Georges and Andersson highlighted the excitement and promise of dissecting quantitative trait loci (QTL) of economic value (Andersson and Georges 2004).

Genome references for the greatest economically impactful food producing species, cattle, sheep, pigs, goats and chickens are now being used to generate large volumes of genotype data that link natural nucleotide variation to phenotypic variation within the context of a production environment. A consequence of access to higher resolution SNP panels and whole genome sequencing (WGS) methods is that QTL are now often resolved to the limits of linkage disequilibrium, even with a keen focus on the more interpretable coding variation. Today some of these loci have been subjected to selection that further advance trait averages with monetary benefits. However, this process is still slow, and beneficial variation can be inadvertently removed or perhaps worse, deleterious variants propagated by linkage. It remains a major challenge to unravel the genes and the regulatory elements that control specific traits before we even consider specific target sequences for genome editing. Should high-value targets be identified, targeted genome editing offers a method to alleviate the disadvantages of selective breeding, mainly the time required to reach a selection target. Despite the advances in genome editing, it is not the sole answer to advance traits of value, but when combined with genomic selection and assisted reproductive technologies, it could transform current livestock improvement strategies.

In his 2005 review of domestic animal genomics, Womack said: “RNA interference may soon find its way into animal improvement, likely in conjunction with cloning from modified somatic cells.” Since this time, genome editing has come of age and been applied to a limited number of food animals (Whitworth *et al*. 2016) (Choi *et al*. 2016) (Carlson *et al*. 2016) (Dimitrov *et al*. 2016). One of the most exciting applications of genome editing is the control of infectious disease, which is a critical need facing livestock producers throughout the world (Smith *et al*. 2016). The host-pathogen relationship have become essential to the spread of new viral strains with major international impact such as new strains of avian influenza on poultry production. To protect future pig populations from devastating viral outbreaks Prather *et al*. edited an entry receptor for porcine reproductive and respiratory syndrome virus infection (Whitworth *et al*. 2016). Another

application of significance to animal welfare and human safety when handling cattle is to generate hornless cattle (Carlson *et al*. 2016). Using knowledge of naturally polled genetic variation (Medugorac *et al*. 2012), the locus responsible was edited to produce hornless cattle thus improving the welfare of cattle by avoiding painful dehorning procedures. These are just a few examples that demonstrate how genome editing can introduce highly valuable natural variants, even those that would be outside of the available breeding population, onto the best genetic backgrounds in one generation without compromising the years of selection of such elite genetic stocks.

The simplicity, scope, and accuracy of genome editing technologies are truly astounding. In fact, our knowledge of the sequences/regions to edit in food producing animals with thousands of QTLs (see http://www.animalgenome.org/cgi-bin/QTLdb/index) already identified for simple monogenic and complex polygenic traits, presents a conundrum as to which targets do we apply this editing capability. An added caveat is that few of these QTLs have definitive causative alleles identified. However, given the economic impact of many of these traits, the incentive to remove or replace associated alleles will eventually lead to genome targets. Of course, this is a simple picture with extensive experimentation required to pinpoint the genes or regulatory regions that will alter the phenotypic outcome. Advances in the characterisation of genes, transcripts and their regulatory regions (a core goal within the FAANG consortium; http://www.faang.org) are likely to underpin the prediction of genes and genetic variant causally linked to simple and complex traits. Genome editing is likely to be an essential tool in our armory to test these predictions in either cell, tissues, organoid or even whole animals. Ultimately, specific genome targets will come into focus and editing experiments will follow. It is expected that equal rigor will be devoted to safety assessments to ensure animal well-being and long-term germplasm diversity, since substantial financial investment will create fewer founders to pass the trait to future generations and, perhaps most importantly, to determine whether the edit meets phenotypic expectations.

It is generally underappreciated that genome editing is just breaking the chromosomal DNA and then allowing the cells natural ability to fix the break precisely and thus incorporate the intended sequence (Segal and Meckler 2013). The basic process is to identify a target sequence to be edited, computationally design a single guide RNA (sgRNA) to introduce the base(s) change, inject the sgRNA and associated reagents into the stem cell, transfer the embryo to the host and if the pregnancy is successful validate the expected edit, and perhaps most importantly start monitoring animal health and performance. The design of sgRNAs has been simplified in the past few years with several bioinformatic pipelines offered (Wong *et al*. 2015) (Doench *et al*. 2016), but if reference errors occur sgRNA design will be flawed and lead to missed targets. Also, to cope with genetic variants and polymorphisms in target genomes, it is necessary to re-sequence many animals in the population and compare them to the reference, again to avoid unwanted off-target sgRNA design errors. Protein-coding gene annotation of food producing genomes is mostly sufficient for sgRNA design to target coding regions. However, paralogs, copy number variants, and non-coding RNAs require further attention in each assembly. Newly available transcript sequencing technology such as Iso-Seq (http://www.pacb.com/applications/rna-sequencing/) will rectify many gene annotation deficiencies (Kuo *et al*., submitted), especially the characterisation of all alternate transcripts and for long non-coding RNA annotation, the most in need of improvement. Also, the functional annotation of animal genomes (FAANG Consortium) will aid annotation of regulatory regions that may be targeted for change once experimental validation catches up.

Most evidence indicates that genome editing, specifically CRISPR methodology, is precise and not off-target (O'Geen *et al*. 2015). However, concerns remain that the edited genome can contain foreign DNA not detected with standard PCR and Southern blot techniques (Kim and Kim 2016). Given the high value of these edited founder animals and the need to ensure a thorough investigation of unexpected off-site effects, we suggest some measures of post-editing genome integrity be implemented. To provide a starting

template for evaluating genome edited food-producing animals, we briefly outline the computational steps using the chicken genome as an example (Figure 1). Our process overview is mostly based on many previously established cancer genome analysis pipelines that compile genetic differences among the genomes of normal and cancer genomes within the patient. Once the genome edited animal is confirmed to contain the targeted base change(s), typically a PCR strategy (Carlson *et al*. 2016), an iterative series of steps is proposed: DNA is extracted from the pre- and post-edited genomes, PCR-free libraries are constructed of short fragment size (~450bp), the genome is sequenced to a minimum of 30x coverage using an X10 Illumina instrument (recommended for cost efficiency) and all sequences (150bp length) are filtered for quality using the PICARD software package module CollectWGSMetrics then mapped using the BWA-MEM aligner to the appropriate animal genome reference for several computational measures. First, any sequences associated with the targeting sgRNA can be identified with fast alignment tools such as BLAT. This step also serves to validate the prior PCR results for base(s) modification. From previous sequence alignments, all single nucleotide polymorphisms (SNPs) and small insertions and deletions (<10bp) are called with two independent callers, such as VarScan2 (Koboldt *et al*. 2013) and Strelka (Saunders *et al*. 2012). Currently the best practice is to converge independent SNP or indel calls to reduce false positives. The converged SNP and indel variant files can be imported into various software tools to evaluate many pre- and post-edited genome properties, for example, we recommend the use of the Ensembl VEP tool (McLaren *et al*. 2010) to catalogue putative loss of function variants within protein coding genes that may impact animal health, although these events could be unrelated to the editing process.

Although it is clear that structural or copy number variants are a major source of variation among humans, their accurate ascertainment is still challenging. The use of physical mapping methods based on whole genome restriction maps is likely to make this easier. We suggest a standard copy number variant analysis, such as CopyCat (Sehn *et al*. 2014), be executed to reveal any significant genome aberrations, i.e. expansions or contractions,

that in some cases can merit further investigation. The tools for this analysis are ever changing, but we offer some choices based on ease of use, accuracy, and sensitivity (Figure 1). Taking advantage of fully developed computational pipelines that generate concise reports of mutation burden in cancer patients will allow these same best practices to be implemented for examining pre- and post-edits to the food-producing genome. Of course, some modifications will be needed. Also, genome editing reports can be modified to account for the regulatory standards that are not clear at this point for food producing animals.

It is exciting to see reference genome assembly completeness and accuracy for many organisms is now nearly reaching quality standards found in the human genome. This development is largely the result of long reads spanning repeats and complete physical maps of chromosomes that allow for *de novo* assembly as never found before. Not surprisingly, we conclude accurate genome assembly and annotation (Not covered here; but an equally important task to define all coding and non-coding transcripts, and their regulatory regions) is required for the success of genome editing experiments. Assuming the genomes of food-producing animals will continue to be edited, we expect standardised methods will be developed and validated to compare genomes before and after genetic manipulation. Measured perturbations to genome integrity or the possibility of finding foreign DNA sequences in animal genomes destined for food consumption compelled us to provide an overview of computational methods and to start discussions of best practices to assure the public that attempts are being made to alleviate concerns about animal welfare or food safety.

ACKNOWLEDGEMENTS

We acknowledge funding from the USDA-ARS 20136701521357 to WCW.

REFERENCES

Andersson L, Georges M. 2004. Domestic-animal genomics: deciphering the genetics of complex traits. *Nat Rev Genet* **5**: 202-212.

Archibald AL, Bolund L, Churcher C, Fredholm M, Groenen MA, Harlizius B, Lee KT, Milan D, Rogers J, Rothschild MF *et al*. 2010. Pig genome sequence--analysis and publication strategy. *BMC Genomics* **11**: 438.

Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM. 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature biotechnology* doi:10.1038/nbt.3238.

Bickhart DM, Benjamin D Rosen, Sergey Koren, Brian L Sayre, Alex R Hastie, Saki Chan, Joyce Lee, Ernest T Lam, Ivan Liachko, Shawn T Sullivan, Joshua N Burton, Heather J Huson, Christy M Kelley, Jana L Hutchison, Yang Zhou, Jiajie Sun, Alessandra Crisa, F. Abel Ponce de Leon, John C Schwartz, John A Hammond, Geoffrey C Waldbieser, Steven G Schroeder, George E Liu, Maitreya J Dunham, Jay Shendure, Tad S Sonstegard, Adam M Phillippy, Curtis P Van Tassell, Timothy P.L. Smith. 2016. Single-molecule sequencing and conformational capture enable de novo mammalian reference genomes. *bioRxiv* doi:http://dx.doi.org/10.1101/063388

: 1-31.

Carlson DF, Lancto CA, Zang B, Kim ES, Walton M, Oldeschulte D, Seabury C, Sonstegard TS, Fahrenkrug SC. 2016. Production of hornless dairy cattle from genome-edited cell lines. *Nature biotechnology* **34**: 479-481.

Chin J. 2014. FALCON: experimental PacBio diploid assembler. https://githubcom/PacificBiosciences/falcon/tree/v013

Choi YJ, Lee K, Park WJ, Kwon DN, Park C, Do JT, Song H, Cho SK, Park KW, Brown AN *et al*. 2016. Partial loss of interleukin 2 receptor gamma function in pigs provides mechanistic insights for the study of human immunodeficiency syndrome. *Oncotarget* doi:10.18632/oncotarget.10812.

Dimitrov L, Pedersen D, Ching KH, Yi H, Collarini EJ, Izquierdo S, van de Lavoir MC, Leighton PA. 2016. Germline Gene Editing in Chickens by Efficient CRISPR-Mediated Homologous Recombination in Primordial Germ Cells. *PLoS One* **11**: e0154303.

Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard R *et al*. 2016. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature biotechnology* **34**: 184-191.

Hastie AR, Dong L, Smith A, Finklestein J, Lam ET, Huo N, Cao H, Kwok PY, Deal KR, Dvorak J *et al*. 2013. Rapid genome mapping in nanochannel arrays for highly complete and accurate de novo sequence assembly of the complex Aegilops tauschii genome. *PLoS One* **8**: e55864.

Kim J, Kim JS. 2016. Bypassing GMO regulations with CRISPR gene editing. *Nature biotechnology* **34**: 1014-1015.

Koboldt DC, Larson DE, Wilson RK. 2013. Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. *Curr Protoc Bioinformatics* **44**: 15 14 11-17.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**: 2069-2070.

Medugorac I, Seichter D, Graf A, Russ I, Blum H, Gopel KH, Rothammer S, Forster M, Krebs S. 2012. Bovine polledness--an autosomal dominant trait with allelic heterogeneity. *PLoS One* **7**: e39477.

Megens HJ, Groenen MA. 2012. Domesticated species form a treasure-trove for molecular characterization of Mendelian traits by exploiting the specific genetic structure of these species in across-breed genome wide association studies. *Heredity (Edinb)* **109**: 1-3.

Myers G. 2014. Efficient local alignment discovery amongst noisy long reads. In *Algorithms in Bioinformatics*, pp. 52-67. Springer.

O'Geen H, Yu AS, Segal DJ. 2015. How specific is CRISPR/Cas9 really? *Curr Opin Chem Biol* **29**: 72-78.

Rubin H. 2011. The early history of tumor virology: Rous, RIF, and RAV. *Proc Natl Acad Sci U S A* **108**: 14389-14396.

Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. 2012. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**: 1811-1817.

Segal DJ, Meckler JF. 2013. Genome engineering at the dawn of the golden age. *Annu Rev Genomics Hum Genet* **14**: 135-158.

Sehn JK, Abel HJ, Duncavage EJ. 2014. Copy number variants in clinical next-generation sequencing data can define the relationship between simultaneous tumors in an individual patient. *Exp Mol Pathol* **97**: 69-73.

Selvaraj S, J RD, Bansal V, Ren B. 2013. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature biotechnology* **31**: 1111-1118.

Smith J, Gheyas A, Burt DW. 2016. Animal genomics and infectious disease resistance in poultry. *Rev Sci Tech* **35**: 105-119.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK *et al*. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**: e112963.

Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, Markovic C, Bouk N, Pruitt KD, Thibaud-Nissen F *et al*. 2016. A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. *G3 (Bethesda)* doi:10.1534/g3.116.035923.

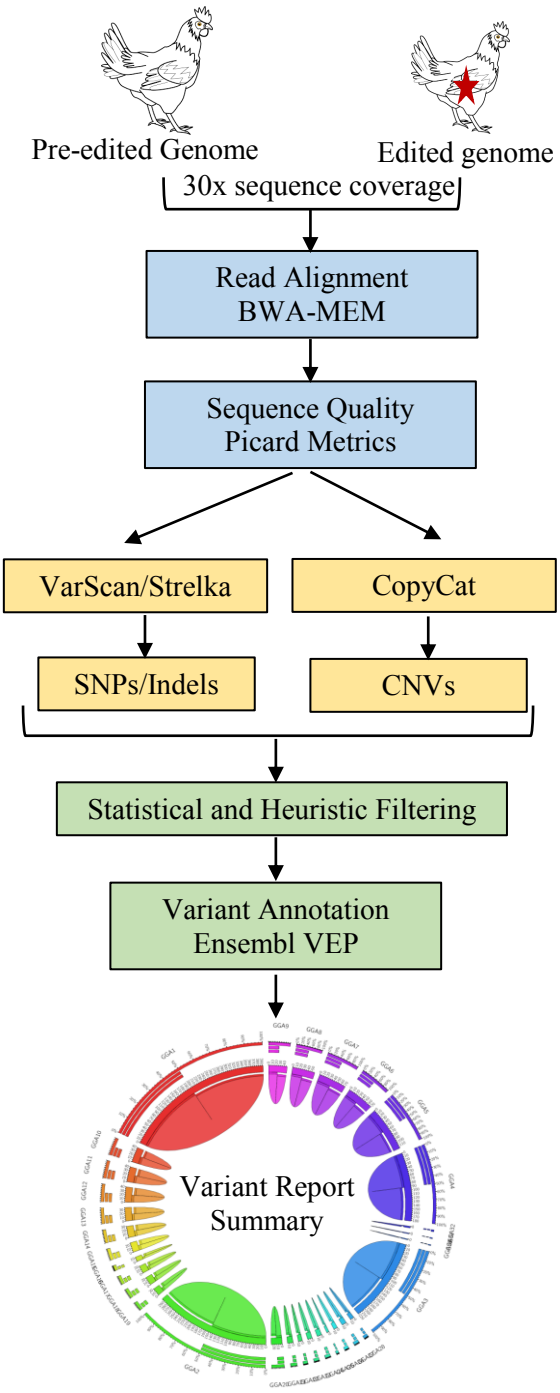
Whitworth KM, Rowland RR, Ewen CL, Trible BR, Kerrigan MA, Cino-Ozuna AG, Samuel MS, Lightner JE, McLaren DG, Mileham AJ *et al*. 2016. Gene-edited pigs are protected from porcine reproductive and respiratory syndrome virus. *Nature biotechnology* **34**: 20-22.

Wong N, Liu W, Wang X. 2015. WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. *Genome Biol* **16**: 218.

Table 1. A summary of food-producing animal genome assembly measures of annotation

Species	NCBI version	N50 contig length	Total contigs	Reference	Protein coding genes	Non-coding genes
Gallus gallus	Gallus_gallus-5.0	2.9 Mb	24,693	(Warren *et al*. 2016)	19,137	6,550
Bos Taurus	Btau_5.0.1	276kb	42,267	None	21,514	5,563
Ovis aries	Oar_v4.0	150kb	48,482	None	20,645	3,861
Sus scrofa	Sscrofa10.2	69kb	243,033	(Archibald *et al*. 2010)	24,205	12,191
Capra hircus	ASM17044v1	26 Mb	30,399	(Bickhart 2016)	20,755	4,011

Figure 1. Computational steps for evaluating genome edited chickens



http://dx.doi.org/10.2218/natlinstbiosci.1.2016.1745