Genes **2012**, *3*, 233-247; doi:10.3390/genes3020233

OPEN ACCESS

genes

ISSN 2073-4425

www.mdpi.com/journal/genes

Article

Case Study of Sequence Capture Enrichment Technology: Identification of Variation Underpinning Developmental Syndromes in an Amniote Model

Elizabeth A. Robb and Mary E. Delany *

Department of Animal Science, University of California, Davis, CA 95616, USA;
E-Mail: earobb@ucdavis.edu

* Author to whom correspondence should be addressed; E-Mail: medelany@ucdavis.edu;
Tel.: +1-530-754-9343; Fax: +1-530-752-0175.

*Received: 6 January 2012; in revised form: 4 March 2012 / Accepted: 5 March 2012 /
Published: 26 March 2012*

Abstract: Chicken developmental mutants are valuable for discovering sequences and pathways controlling amniote development. Herein we applied the advanced technologies of targeted sequence genomic capture enrichment and next-generation sequencing to discover the causative element for three inherited mutations affecting craniofacial, limb and/or organ development. Since the mutations (*coloboma*, *diplopodia-1* and *wingless-2*) were bred into a congenic line series and previously mapped to different chromosomes, each targeted mutant causative region could be compared to that of the other two congenic partners, thereby providing internal controls on a single array. Of the ~73 million 50-bp sequence reads, ~76% were specific to the enriched targeted regions with an average target coverage of 132-fold. Analysis of the three targeted regions (2.06 Mb combined) identified line-specific single nucleotide polymorphism (SNPs) and micro (1–3 nt) indels. Sequence content for regions indicated as gaps in the reference genome was generated, thus contributing to its refinement. Additionally, Mauve alignments were constructed and indicated putative chromosomal rearrangements. This is the first report of targeted capture array technology in an avian species, the chicken, an important vertebrate model; the work highlights the utility of employing advanced technologies in an organism with only a "draft stage" reference genome sequence.

Keywords: capture array; next generation sequencing; bioinformatics; chicken; mutation; SNP; variant; indel; congenic

1. Introduction

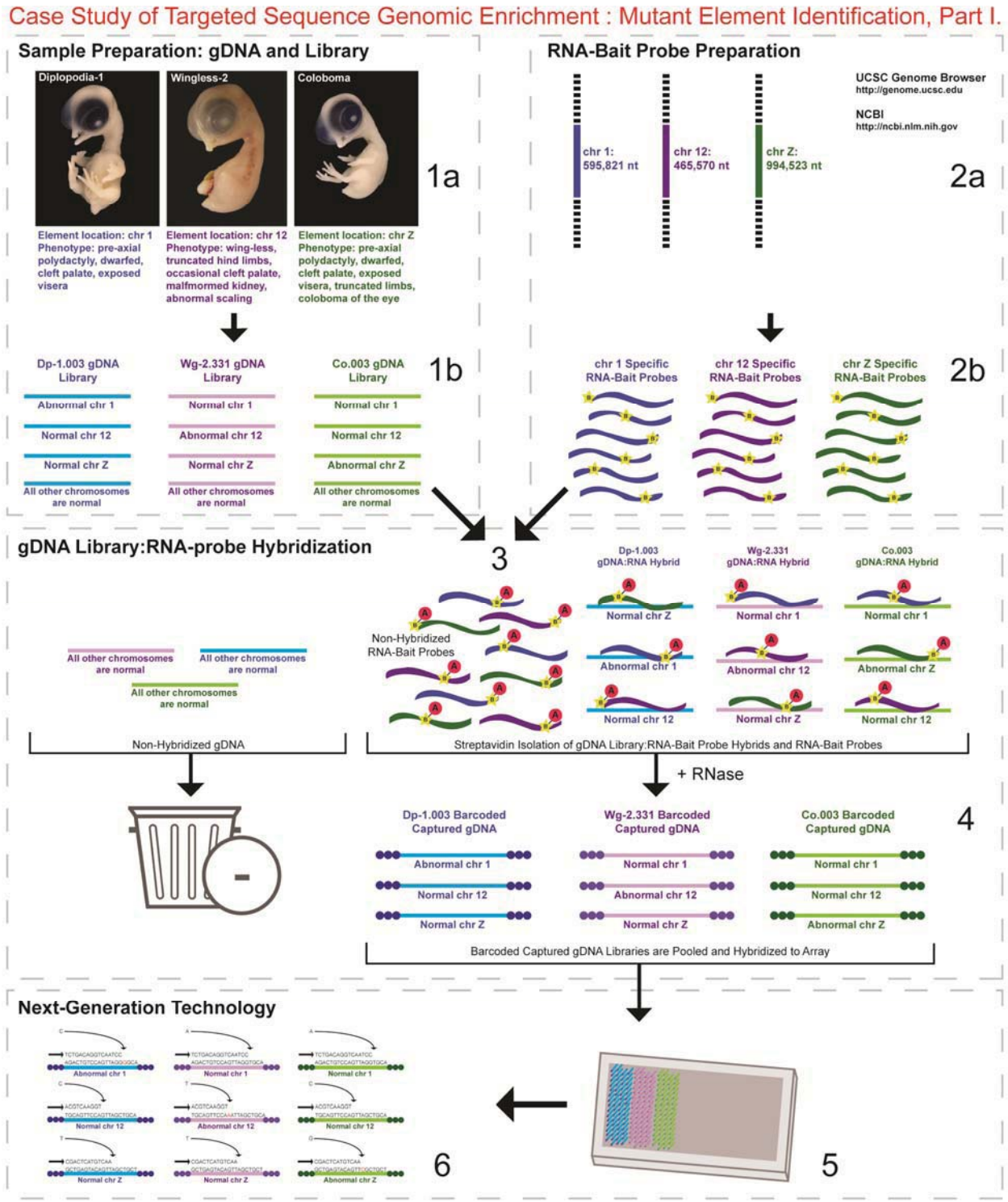
Next-generation sequencing (NGS) enables researchers to accurately and rapidly address many critical and outstanding biological questions regarding the relationship between genotype and phenotype. Employment of a targeted genomic enrichment capture array (CA) approach can advance such research contingent on sequence knowledge of the genome of interest or a closely related genome. Paired with comparative developmental biology, NGS can elucidate the sequences responsible for vertebrate developmental malformations [1]. Like many laboratories worldwide, we were interested in employing the latest technologies, in a cost-effective manner, to understand a biological system despite lacking substantial resources or expertise in bioinformatics. Our genome of interest, the chicken, has not been mapped to the precision, detail, or accuracy of that of the mouse or human [2–4]. To this end, we employed a commercial service provider and made optimal use of specialized genetic resources to identify sequence variation associated with mutations for the long-term objective of determining the causative elements and their role in amniote developmental pathways.

The University of California-Davis (UCD) maintains a series of developmental mutant chicken lines well-studied for phenotype and mode of inheritance [1,5–7]. The mutations are single-gene recessives causing craniofacial, limb, skeletal, muscular and/or integument abnormalities having homology with developmental syndromes in human. The mutations were bred on the same inbred genetic background (UCD-003) to generate congenic (a.k.a. coisogenic) inbred lines, thereby providing an advantage for discovery of the specific genetic element causing each defect. The chromosomal locations and causative regions (CR; a.k.a. linked region) associated with each of the three mutations were previously mapped using SNP genotyping arrays (*diplopodia-1* (*dp-1*) mapped to GGA 1; *wingless-2* (*wg-2*) mapped to GGA 12; *coloboma* (*co*) mapped to GGA Z, sex-linked) [1,8].

Here, we utilized capture enrichment technology rather than exome or whole genome sequencing (WGS) for several reasons, including: (1) variability in the phenotypic expression of each syndrome suggested that the causative element could reside within a regulatory element rather than exon; (2) the reference chicken genome sequence is still in an early stage (gene annotation not as robust) thereby possibly leading to missed genes; (3) we have the advantage of congenic lines and knowledge of specific regions associated with each mutation, therein WGS was unnecessary; and (4) at the time, the cost for WGS three mutant lines was higher compared to the capture array technology. Therefore, we targeted the specific genomic coordinates of the three CRs on GGA 1, 12, and Z, totaling 2.06 Mb, to sequence, in their entirety, those regions known to maintain the elements causing the three unique developmental mutations. The CA technology consists of a "targeted-genomic" aspect wherein overlapping oligonucleotide RNA-bait probes are generated for a specific genomic segment (in our case, maximum CR linked to each mutation). The probes are hybridized to DNA from the target (in our case, three developmental mutants barcoded for line-specificity), amplified and sequenced using NGS methods in the "capture enrichment" aspect of the technology (Figures 1 and 2). The bioinformatics

analyses identified genetic differences (e.g., SNPs, micro-indels, and macro-indels) providing information on both normal variation from the introgressed region(s), as well as priority mutant-specific sequences for future functional studies.

Figure 1. Targeted Sequence Genomic Enrichment Methodology, Part I. Targeted genomic capture enrichment paired with next-generation sequencing technology was utilized to sequence, in their entirety, the three chromosomal regions associated with three developmental mutations in the chicken. Approximate timeline from contract with the service provider to transfer of data (steps 1–6) was ~6 months.



1a. *Identified regions for targeted sequence capture enrichment.* The linked/causative regions were previously identified using SNP arrays [1]. High quality, high molecular weight genomic (g) DNA was isolated from specific samples of interest (Agilent SureSelect kit suggestions: 10–20 µg of purified, non-amplified gDNA per sample with concentrations between 100–1,000 ng/µL, A_{260}/A_{280} ratio >1.8). The samples were assessed for quality (e.g., spectrophotometric analysis, agarose gel quality control). The quality of the gDNA and library (e.g., library size distribution and concentration) generated will have an impact on the quality of the sequence results.

1b. *Prepared genomic libraries.* Three individual libraries were generated from gDNA isolated from three developmental mutant congenic lines using Agilent's SureSelect Target Enrichment System. The following steps were taken in the preparation of each individual mutant library: (i) shear DNA to obtain fragments with a base pair peak of 150 to 200; (ii) blunt-end fragments with 5′-phosphorylated ends; (iii) attach a dATP to the 3′ end of the DNA fragments. After dATP nucleotides are added to the 3′ end of the DNA fragments; (iv) adaptors (specific to the sequencing platform) are ligated to the 3′ dATP overhang; (v) a library pre-enrichment amplification followed by (vi) a library quality control and quantitation assessment with a Bioanalyzer and PicoGreen assay. Please note that a purification procedure occurs in between each of the library preparation steps (i–vi). If the initial or enriched template library contains low amounts of nucleic acid, one can amplify the library before sequencing using PCR and a polymerase that is not biased as to template size. One can outsource any of the subsequent steps or perform them in the research laboratory.

2a. *Designed overlapping RNA-bait probes (120 nt in length) for sequencing for the region of interest.* Sequence information for chromosomes 1 and 12 was obtained from NCBI (WASHUC2, May 2006 [2]) while red jungle fowl, UCD-001 (reference genome genetic line), sequence data for chromosome Z was obtained from Dr. D. Winston Bellott and Dr. David Page prior to NCBI submission [9]. We provided SeqWright Inc., with coordinate information or sequence content to design and create the RNA-bait probes that complemented the three targeted chromosomal regions of interest.

2b. *Generated overlapping biotinylated RNA-bait probes specific to each region of interest*. RNA library "baits" were generated for bead capture purposes (step 3).

3. *Hybridized denatured gDNA library fragments (150–200 nt) to RNA-bait probes (120 nt).* RNA-baits were hybridized to gDNA in order to enrich for complimentary DNA sequence information specific to the three regions of interest. Streptavidin coated magnetic beads were utilized to capture RNA-bait:gDNA-library fragment hybrids as a means to separate those DNA fragments not complementary to the targeted regions. Beads were washed and digested (RNased) to isolate only gDNA library fragments that hybridized to RNA-bait probes.

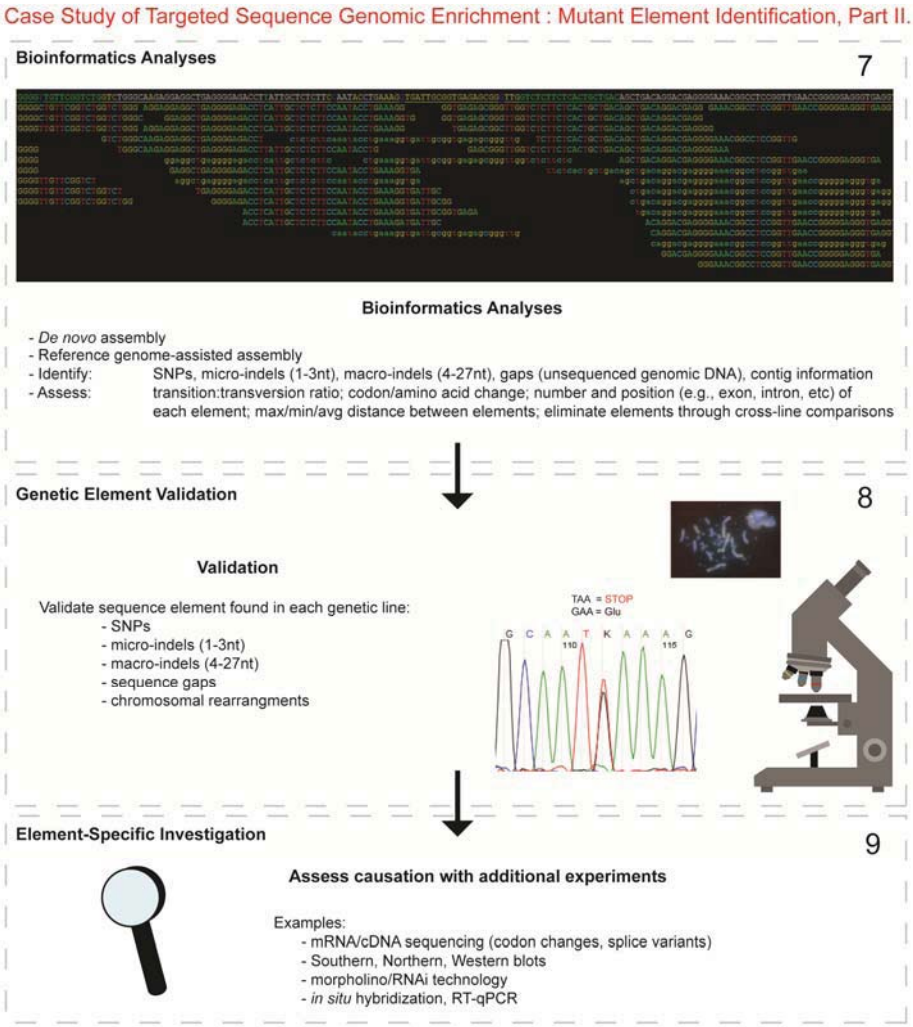
4. *Barcoded samples (pooled or individual).* In order to identify each individual or group, samples were barcoded (a.k.a. index-tagged). In the case of this project, Co.003 gDNA (2 pooled female mutants), Dp-1.003 (2 pooled female mutants), and Wg-2.331 (2 pooled female mutants), each had an individual barcode unique to the genetic line. [Prior to sequence read alignment and bioinformatic analysis, the barcoded sequencing reads were first sorted and the barcode was then removed.]

5. *Pooled barcoded samples, hybridized sequences to array, and amplified DNA prior to sequencing.* All three genetic lines were pooled into one sample with each barcoded sample

present in equimolar amounts. Pooled, barcoded libraries (single-stranded) were hybridized to an array (in the case of this project, we used one-fourth of a slide) utilizing the adaptors (see step iv in 1b) previously incorporated at the end of the DNA sequence. Unlabeled nucleotides and enzyme were added to initiate solid-phase bridge amplification (this generates double-stranded bridge molecules), DNA was then denatured, and amplification to generate sequence clusters proceeded.

6. *Next-generation sequencing.* Labeled dNTP reversible terminators (one base at a time), primers and DNA polymerase were added to the slide and sequenced using laser excitation. This step was repeated until each barcoded DNA fragment was sequenced. For this project, SOLiD™ version 3-Plus using 50 ligation cycles (50 base pair sequencing) was employed. A total of 3.64 Gbp of sequence data was generated.

Figure 2. Targeted Sequence Genomic Enrichment Methodology, Part II. Analysis of the targeted genomic capture enrichment and next-generation sequencing data allowed for the identification of variants and chromosomal rearrangements which were further validated using new mutant samples in order to identify the causative element for each of the developmental mutations. Approximate timeline from obtaining the raw SOLiD™ (colorspace) sequence reads to validating the sequence variants identified (steps 7–8) was 6 months.



7. *Bioinformatics.* We received colorspace reads and quality value files (both in FASTA-like formats) from SeqWright, Inc. SNPs, micro-indels (1–3 nt), macro-indels (4–27 nt) and gaps were identified for each of the three genetic lines. Several reference-assisted *de novo* assemblies were generated using Mauve 2.3.1c software [10] in order to identify chromosomal rearrangements. Additional variant analyses included, but were not limited to, identification of: transition:transversion ratios, codon and amino acid modifications due to presence of the element under study, number of each element and maximum/minimum/average distance between variants, as well as position within the genome (e.g., exon, intron, splice site, *etc*.).
8. *Validated genetic elements*. Variations were assessed for linkage to the mutant and to evaluate the contribution of the polymorphisms to the mutant phenotype and/or genetic line using a new cohort of individuals ($n \geq 20$) [11].
9. *Element-specific investigation*. Each genetic element must be further assessed to discover causation towards the phenotype. Examples of additional assays include: mRNA/cDNA sequencing (codon/amino acid modifications, splice variants), Southern, northern, western blots, morpholino/RNAi, *in situ* hybridization, RT-qPCR, and/or chromatin studies.

2. Results and Discussion

The development of NGS technologies is advantageous for all genomics research, especially for those genomes not as well annotated as mouse and human. Such technologies are of interest to laboratories of varied size, scope, and funding resources. NGS is no longer limited to novel sequence identification, but also offers insight into sequence variation, quantitative gene expression analysis, and questions of evolutionary genomics. Applications of the CA technology include but are not limited to re-sequencing of exonic regions, candidate gene sets, and large genomic loci as well as biomarker discovery and genetic marker development. Further, CA can provide a better understanding of the genetic basis for polygenic diseases, metabolic pathways, and in the case of this study, sequencing of unique genetic lines. To date, CA studies have been reported in mouse and human for the aforementioned applications, as well as a few domesticated species with many more plant and animal systems in the pipeline [12–21]. CA studies have led to the discovery of more than 40 Mendelian disorders in human [22].

Here, we describe the utilization of targeted sequence genomic enrichment in an important vertebrate model with a relatively early-stage reference genome [2], taking advantage of congenic inbred lines. Our long-term objective is to utilize the information gained to establish the specific element or gene responsible for each of the inherited recessive mutations known to result in abnormal developmental phenotypes [1]. This technology was successful in identifying numerous sequence elements unique to each genetic line in addition to reducing the size of the linked region (Tables 1 and 2), thereby eliminating some candidate genes. Figures 1 and 2 present a schematic, displaying the methodology utilized in our study, and provides a step-wise path for other researchers. Figure 1 begins with the collection and preparation of samples and the genomic libraries (step 1a, b), describes the design and generation of RNA-bait probes (step 2a, b) and the subsequent hybridization of the probes to the individual libraries (step 3) for genomic enrichment. The enriched libraries are then index-tagged

(for identification purposes) (step 4) and sequenced (step 5, 6). Figure 2 describes the bioinformatics analyses conducted (step 7) and discusses additional functional assays (steps 8, 9).

Table 1. Statistics of the three regions captured using the targeted genomic capture enrichment technology.

Genetic Lines	Total Sequenced Region					
	Chr	Size (nt)	No. Genes	No. Gaps [A]	NCBI Mean Quality Score	GC richness
Diplopodia-1.003	1	595,821	19	5	87.3%	39.6%
Wingless-2.331	12	465,570	13	15	94.5%	41.7%
Coloboma.003	Z	994,523	6	0	94.5%	39.4%
Total	-	2,055,914	38	20	-	-
Average	-	685,305	12.7	7	92.1%	40.2%

Three unique chromosomal regions were targeted for utilization in the genomic capture enrichment technology. Descriptive measures include targeted chromosome and size, number of genes and sequence gaps found within targeted region, percent GC richness and quality score. [A] Gaps were identified through assessment of each linked, genomic region for the three mutations using the UCSC Genome Browser [4]. Gaps identified for the Dp-1.003 chromosome 1 region were ≤1000 nt while Wg-2.331 region gaps were ≤1500 nt. No gaps were present in the 995 kb chromosome Z region for the Co.003 genetic line. However, given the repetitive nature of the Z chromosome, probes were only generated for 990,270 of the 994,523 nts. Note that the 4253 nts were present in a genomic region shown to no longer be linked to the *Co*.003 mutation [1].

2.1. Three Region (2.06 Mb) Sequencing and Mapping

SeqWright's Genomic Enrichment Services were employed to design a custom NimbleGen capture array to enrich for sequencing the chromosomal regions linked to three developmental mutations in the chicken. Using only one-fourth of a slide and SOLiD ™ version 3-Plus using 50 ligation cycles (50 bp sequencing) a total of 3.64 Gbp of sequence data were generated. Using a short read mapping program, BWA, 72.6% of all reads were mapped to the three linked chromosomal segments; these mapped reads covered 1.95 Mb of the combined 2.06 Mb probe-generated regions (Table 2).

2.2. Analysis of Three Congenic Developmental Mutant Genomes

The following elements were assessed from the comparisons among the three genetic mutant lines using several programs and custom scripts (see Methods): SNPs, insertions (1–3 nt), deletions (1–3 nt), and gaps (unsequenced regions). A total of 2593, 1724, and 2500 SNPs were found within the three mutant CRs (GGA 1, 12, and Z) for dp-1.003, wg-2.331, and co.003, respectively. The insertions and deletions were 150 and 133, respectively, for dp-1.003 on GGA 1, 108 insertions and 138 deletions for wg-2.331 (GGA 12), and for co.003, a total of 125 insertions and 155 deletions were identify within the CR on GGA Z (Table 3). Only those variants localized within the 2.06 Mb targeted region are reported, all others are considered non-specific.

Table 2. Summary of targeted genomic enrichment sequencing results for three developmental mutant congenic lines.

Genetic Lines	Sequencing Read Statistics					Region Reduction (Post-Analysis)		
	Total Reads Generated [A]	Total Mapped Reads [B]	Average Coverage [C]	Region Sequenced [D]	No. Coverage Gaps [E]	Remaining Size (nt)	Fold Reduction	No. Genes Remaining
Diplopodia-1.003	21.0M	15.5M	107.2×	96.9%	232	261,947	2.3×	12
Wingless-2.331	36.0M	28.3M	217.1×	85.3%	274	259,545	1.8×	13
Coloboma.003	15.7M	11.9M	72.1×	98.4%	525	306,847	1.3×	5
Total	72.7M	55.7M	-	-	1,031	828,339	-	30
Average	24.2M	18.6M	132.1×	93.5%	344	276,113	1.8×	10

Targeted genomic capture enrichment results: read statistics and post-analysis assessment. Descriptive measures include: total reads generated, total reads mapped to targeted region (2.06 Mb), average fold coverage, percentage region sequenced, and number of sequence/coverage gaps. Assessment of the sequence reads resulted in a reduction in the linked-region size. [A] Total number of reads generated, M = million; [B] Total reads mapped to the 2.06 Mb sequence used in the capture array, M = million. The entire chicken genome (*Gallus gallus* v2.1 (galGal3) assembly (WASHUC2, May 2006)) was used in the mapping (chr 1–28, W, and the 995 kb Z) rather than only the three linked chromosomal segments; [C] Average fold sequence coverage of the 2.06 Mb targeted region; [D] Region sequenced refers to the dp-1, wg-2, or co (596, 466, and 995 kb, respectively) sequence information in which RNA-bait probes were generated from and sequence data aligned to. For example, of the 595 kb dp-1 sequence information for which RNA-bait probes were generated, 96.9% or ~577 kb had sequence reads map to it. Thus 3.1% of the region had no reads mapped to it; [E] A gap was defined as any fragment of DNA absent in a genetic line relative to the reference genome sequence. A gap could be due to: (1) an RNA-bait probe was not designed correctly; (2) the NGS technology failed to sequence the fragment of DNA due to sequence structure/quality (e.g., repeat, GC-rich); (3) the reference genome was composed of unknown sequence (N) and therefore a probe could not be generated for that region but the size of the fragment was still known and accounted for in the reference genome; and (4) the genetic line of interest does not have that fragment of DNA, *i.e.*, the gap is a "true deletion".

Table 3. Assessment of SNPs and micro-indels within three congenic lines: Diplopodia-1.003, Wingless-2.331, Coloboma.003.

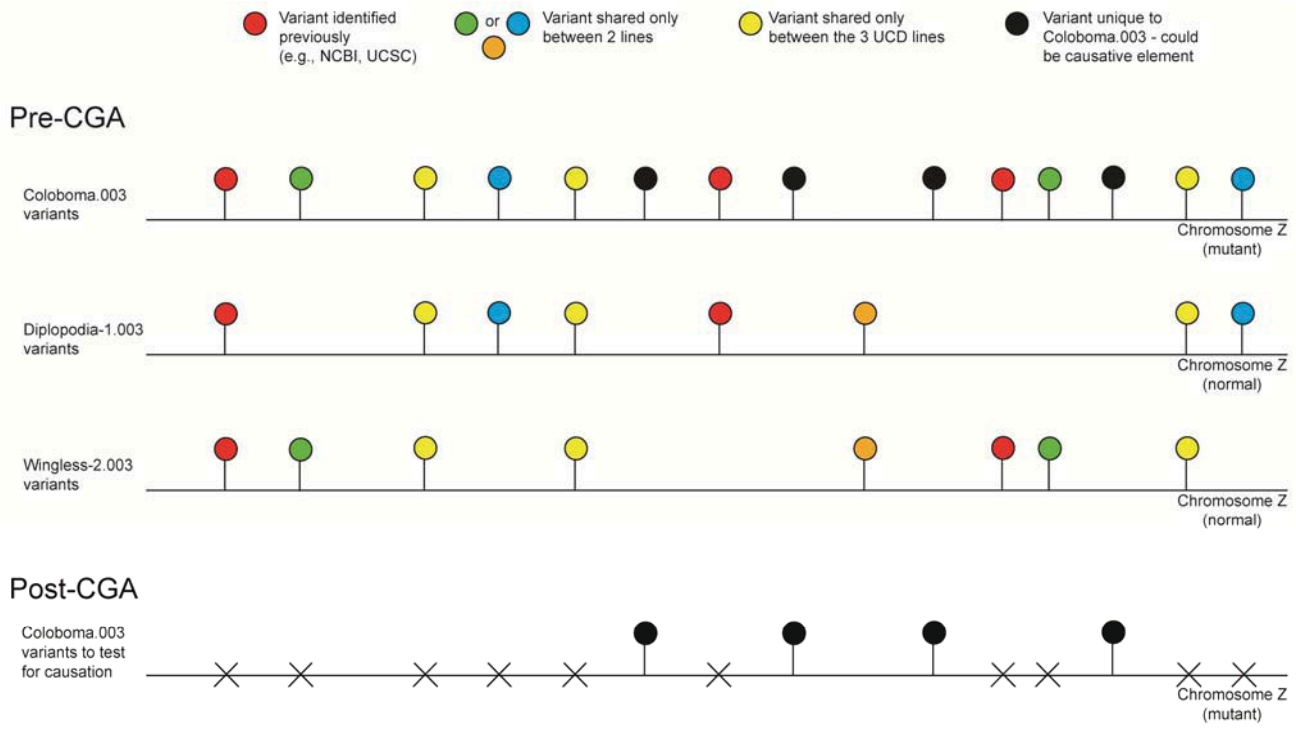
Chr	# of SNPs in Sequenced Region			# of Short Indels (1–3 nt) in Sequenced (2.06 Mb) Region					
	Dp-1.003	Wg-2.331	Co.003	Dp-1.003		Wg-2.331		Co.003	
				Insertions	Deletions	Insertions	Deletions	Insertions	Deletions
1	***2,593***	2,434	2,478	***150***	***133***	116	130	109	110
12	1,245	***1,724***	1,225	79	101	***108***	***138***	71	93
Z	2,903	1,787	***2,500***	128	171	150	185	***125***	***155***

Total number of SNPs, insertions, and deletions identified within the targeted regions of interest.

Multiple pairwise-line comparisons (a.k.a. comparative genomic analysis) were conducted to eliminate SNPs and indels shared between two or more of the congenic lines or to previously identified polymorphisms found in databases (see Figure 3). We identified a total of 6104 SNPs, 245 insertions, and 299 deletions shared among the congenic lines within the linked regions of GGA 1, 12, and Z; these results are suggestive of polymorphisms found within the highly inbred UCD-003 genetic line and will be useful in creating future SNP arrays (NCBI accessions: ss472336609-ss472343089). The pairwise-line comparison indicated 2110 novel SNPs, 201 novel insertions, and 219 novel deletions unique to the three genetic lines, *i.e.*, the respective introgressed regions. Assessment of homozygous and heterozygous SNP loci reduced the size of the linked regions by 1.23 Mb in total, with an average of

276 kb remaining as linked to each mutation (Tables 1 and 2). Further, assessment of the SNPs alone allowed us to eliminate eight genes as causative for the mutant phenotypes, a major advantage for future functional studies. Similarly, micro-indels (1–3 nt) and gaps were identified (Table 2) and reduced in the same fashion by multiple pairwise-line comparison. These variants will be assessed for contribution towards the mutant phenotype. Macro-indels (4–27 nt) were also identified; however, upon review of those positioned within the exons and splice sites of Wg-2.331 only, none of the bioinformatically predicted macro-indels (n = 16) were present (data not shown). The 0% validation rate for the macro-indels indicates a custom script problem; this script will be revised and macro-indel location re-evaluated. Further assessment of the unique sequence elements identified (SNPs, micro-indels, gaps) include, but are not limited to, identification of the position of each element in the genome (e.g., exon, intron, *etc*.) and codon and amino acid modifications caused by the element (Figure 2, step 7). Such knowledge allows prioritization of the elements for further validation (Figure 2, step 8–9).

Figure 3. Illustration of the comparative genomic analysis strategy to identify mutant-specific polymorphisms. Comparative genomic analyses (CGA) were conducted to eliminate SNPs and indels shared between two or more of the congenic lines or to previously identified polymorphisms. Those variants not observed in any other genetic line or chicken species were denoted as unique and potential causative elements. An example, to illustrate how this analysis was performed, is shown in reference to the UCD-Coloboma.003 linked region on chromosome Z. All three genetic lines were assessed at the same location for polymorphisms, those variants shared between Co.003 and any other line were removed. Only those elements unique to that region in the Coloboma.003 genetic line are of interest and should be further assessed for causation.



2.3. Reference-Assisted De Novo *Assembly*

Genomic alignment programs are utilized to: (1) identify putative inversions and other genomic rearrangements; (2) confirm the current reference genome assembly; (3) assess gaps in the sequence information (a gap can be present due to sequencing problems or because the gap is a deletion in a particular line); and (4) identify novel genomic sequence. Reference-assisted *de novo* assembly was therefore conducted using the Mauve 2.3.1c software [10] on the 72.7M reads generated for Dp-1.003, Wg-2.331, and Co.003 genetic lines (Table 2). Seven independent Mauve alignments were conducted to identify contig-alignment program bias. This included four three-chromosome alignments (Supplemental Figure 1), utilizing only the linked regions of GGA 1, 12, and Z, and three individual chromosome assemblies (Supplemental Figure 2). From these reads Mauve generated an average of 81 contigs ≥400 bp for each of the three-chromosome assemblies; 89.4% of these contigs showed high sequence similarity (≥90%) to the reference genome. The program results indicated several translocation events. Validation of the chromosomal rearrangements and identified gaps using a variety of techniques (e.g., FISH, re-sequencing) is necessary in order to assess legitimacy of any predicted chromosomal rearrangement.

3. Experimental Section

3.1. Genetic Lines

We investigated three developmental mutations bred into congenic lines: Diplopodia-1 (Dp-1).003, Wingless-2 (Wg-2).331, and Coloboma (Co).003. Two of the single-gene mutations (*dp-1*, *co*) were backcrossed onto the highly inbred (F >0.99) Single Comb White Leghorn (SCWL) UCD-003 and the third (*wg-2*) was bred to be congenic on UCD-331, which is a congenic line on the UCD-003 background except for the major histocompatibility complex [1,23]. The congenic line series provides a unique advantage for employing capture array technology and subsequent analyses as each line serves as a control genotype for the other two since the mutants map to three different chromosomal regions. Phenotypes, SNP-genotyping, and mapping of the mutant chromosomes and CRs are described by Robb *et al*. [1]. Animals used for the study were under the care and supervision of trained staff and as per protocols approved by the UC Davis Institutional Animal Care and Use Committee.

3.2. Sample Collection, Capture Array Sample Preparation, and SOLiD ™ Sequencing

Embryos were incubated to E10, an age of development when the phenotypes are easily discerned [normal +/+ and +/−; mutant −/− (autosomal) or −/W (sex-linked)]. DNA was obtained from three tissues (brain, heart, and liver) and purified using the DNeasy® Blood & Tissue kit (Qiagen). Mutant status was also confirmed by genotyping using primers linked to each mutation [1].

As *co* is sex-linked, we used females which increased the probability that the mutant Z region would be sequenced (in birds, the female is the heterogametic sex (ZW) while the male is the homogametic sex (ZZ)). Thus, male (ZZ)/female (ZW) status was determined based on phenotype (gonadal examination) and genotype. The latter employed standard PCR of two sex-chromosome-specific loci [1]. Reactions were amplified using Phire® Hot Start II DNA Polymerase (Finnzymes) and amplicons sized by

electrophoresis (1.5% gel, 1× TAE, 100 V, 2 h). Primers were GGA_W_F (5′ -CTGACTACCTT TGCAGTGCT- 3′), GGA_W_R (5′ -GCTGAGAAACTTATCCCTCA- 3′), GGA_Z_F (5′ -AAAGCA AAGGTTTTTGTTCC- 3′), and GGA_Z_R (5′ -TGGAAATGCCTGCTAAACTA- 3′). Amplicon sizes were 227 bp (GGA W) and 161 bp (GGA Z).

Line-specific DNA pools (50 μg/line of two female (25 μg each) mutant samples/line), were sent to SeqWright DNA Technology Services (Houston, TX) for target enrichment and SOLiD ™ V3 Plus Platform sequencing (Applied Biosystems, Foster City, CA). Only 3 μg of DNA is needed to generate a library so the amount sent was in large excess. SeqWright's Genomic Enrichment Services utilized Agilent's (Santa Clara, CA) SureSelect Target Enrichment System in order to enrich for three specific regions on chromosomes 1 (Dp-1.003), 12 (Wg-2.331) and Z (Co.003) using overlapping RNA sequence "baits" (120-bp in length, complementary to the targeted region) selected specifically to provide complete coverage for each genomic region of interest [1]. GGA 1 and 12 reference sequences for probe design were obtained from NCBI [24] and GGA Z sequence information from Drs. D. Winston Bellott and David Page (Whitehead Inst., MIT) [9]. Preceding target enrichment and sequencing, a fragment library (base pair peak of 150–200) was constructed for each genetic line. Each library was then tagged with a different barcode (a.k.a. indexing-specific adapter) so that the individual genetic lines were distinguishable; the three libraries were pooled prior to emulsion PCR and SOLiD™ sequencing. See Figures 1 and 2 for more detail as to the CA/NGS methodology.

3.3. Sequence Assembly

Colorspace read data (.csfasta, .stats, .qual) were sorted by the barcodes using custom Python scripts; barcodes were removed and reads (50 bp length) were trimmed by three nucleotides on each end. The reads were then mapped to the chicken reference genome (*Gallus gallus* v2.1 (galGal3) assembly (WASHUC2, May 2006) [25]) using Bowtie 0.12.5 [26], allowing for ≤2 mismatches within each read for initial assessment of alignment quality. All alignments were converted to Sequence Alignment/Map (.sam) format and were viewed using Samtools 0.1.7a [27]. Bioinformatic raw data and mapped alignments were outsourced to the University of California, Davis (UCD) Bioinformatics Core [28] for variant identification and reference-assisted *de novo* assembly.

3.4. Reference-Assisted De Novo *Assembly*

Mauve 2.3.1c, a multiple genome alignment software [10], was utilized to identify potential chromosomal rearrangements; these events were identified by analyzing the position of the Locally Collinear Blocks (LCBs) within each genetic line relative to each other. The program first utilizes the original, sorted sequence reads from the three mutant genomes to generate contigs *de novo*. Upon completion, each contig is aligned to the reference genome, thereby identifying its position in the genome. Herein this alignment is referred to as the reference-assisted *de novo* assembly.

3.5. SNP, Micro- and Macro-Indel, and Sequence Gap Discovery

The trimmed read data and alignment files were provided to the UC Davis Bioinformatics Core wherein read format conversion and alignment using BWA [29,30], SAMtools [27,31], and custom Perl

scripts were run. Output files included information on SNPs, micro-indels (1–3 nt) and macro-indels (4–27 nt) for all three mutant regions. Filters for the identification of SNPs and micro-indels include: read coverage ≥10, root-mean-square (RMS) mapping quality of the aligned reads ≥30, and the minimum variant frequency/count ≥20% (*i.e.*, ≥2 mutant read variants per 10 reads). Note that SNP and micro-indel identification was carried out for each genetic line individually, as reads were sorted by barcode prior to alignment and variant discovery. Additionally, gaps (unsequenced segments/coverage gaps) within the 2.06 Mb targeted region were identified using BEDTools [32,33]. Those genomic gaps only identified in one line (but present in the others) were considered putative deletions and should be confirmed for legitimacy.

To evaluate the accuracy of variant calling, SNPs identified as linked to each mutation using the 3K and 60K SNP arrays [1] were identified as present in the CA results. Additionally, both SNPs and micro-indels (1–3 nt) were compared to previously identified variants found in other chickens (*i.e.*, Silkie, commercial egg- and meat-type birds assessed by the Beijing Genomics Institute, [34]) [NCBI, UCSC Genome Browser].

3.6. Capture Enrichment Data Analyses

Sequences were initially identified as polymorphic relative to the *Gallus gallus* NCBI reference genome sequence, UCD-001 (Red Jungle Fowl (RJF)) [2]. However, since such differences would include normal variation (unrelated to the mutation due to line divergence between the RJF and SCWL), we reduced this variation by employing a comparative genomic analysis between and among the three mutant lines (Figure 3). Previously, we had shown that the UCD developmental mutant lines varied at only the causative region (CR) while the remaining genome was largely identical to UCD-003 [1]; consequently, each mutant line serves as a control/reference sequence for the other two mutations since the mutations are linked to different chromosomes.

Thus, our strategy for mutant-specific sequence discovery was the following: after sequence variant identification and genome comparisons, any polymorphism identified as unique, defined as present only in the mutant line of interest and not found in the other two lines, the reference genome, or any previously identified polymorphism reported in NCBI and the UCSC Genome Browser, was considered a candidate for being the causative element for the respective developmental mutation (Figure 3).

4. Conclusions

With the price for NGS technologies decreasing at exponential rates, this state-of-the-art technology is now a realistic research tool for many biological laboratories desiring genomic sequence information to complement biological analysis of phenotypes. However, two limiting factors still exist, thereby discouraging the use of NGS in smaller labs. Such deterrents include the lack of bioinformatics expertise and a "draft-stage" reference genome. To overcome these two issues we: (1) outsourced our sequencing data to an affordable, local bioinformatics core and received results within a reasonable period of time and (2) utilized not only the closely related species reference genome (*Gallus gallus versus G. domesticus*) but also congenic line partners for alignment, variant discovery and elimination. Beyond contributing to our analysis of the mutations, the information generated contributes to an

improved understanding of the reference genome, inclusive of the filling-in of sequence gaps and basic information of sequence content of particular chromosomes (e.g., autosomal *vs.* sex) (unpublished data).

In summary, the utilization of the CA/NGS technology resulted in the narrowing of the causative region size for each of the three mutations studied. Further, our multiple pairwise-line comparison strategy eliminated those shared polymorphisms unrelated to the syndromes. Sequencing verification of the identified SNP, insertions, and deletions using additional mutant samples will indicate if a particular polymorphism remains linked, *i.e.*, a validation step [11]. Ultimately, to establish the role of a sequence element in causing the mutant phenotype, such sequence variants will be analyzed using the appropriate functional assays to determine cause/effect.

Acknowledgments

We are grateful to D. Winston Bellott and David Page (Whitehead Inst., MIT) for the UCD-001 GGA Z sequences. We thank Clayton Morrison and Fei Lu, and Adrian Costa at SeqWright, Inc. for their expert assistance and advice with the capture array development and technology. We thank C. Titus Brown for his advice and hosting of the 1st Annual Michigan State University Bioinformatics course attended by Elizabeth A. Robb and the UC Davis Bioinformatics Core for their expertise. We acknowledge the funding sources: USDA-NIFA Multi State Research Project NRSP-8, NC-1170, and the UC Davis John and Joan Fiddyment Endowment. Infrastructure and genetic line resources were supported by the California Agricultural Experiment Station, the Department of Animal Science, and the College of Agricultural and Environmental Sciences at the University of California, Davis. The authors declare no competing interests.

References

1. Robb, E.A.; Gitter, C.L.; Cheng, H.H.; Delany, M.E. Chromosomal mapping and candidate gene discovery of chicken developmental mutants and genome-wide variation analysis of MHC concgenics. *J. Hered.* **2011**, *102*, 141–156.
2. International Chicken Genome Sequencing Consortium (ICGSC). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **2004**, *432*, 695–716.
3. Dodgson, J.B.; Cheng, H.H.; Warren, W.C.; Zimin, A.V. Proposal to Enhance the Sequence of the Chicken Genome—2010 Whitepaper. Available online: http://poultry.mph.msu.edu/about/Chicken_Genome%20supplemental%20sequencing.pdf (accessed on 5 October 2011).
4. UCSC Genome Bioinformatics Homepage. Available online: http://genome.ucsc.edu (accessed on 13 March 2012).
5. Pisenti, J.M.; Delany, M.E.; Taylor, R.L.; Abbott, U.K.; Abplanalp, H.; Arthur, J.A.; Bakst, M.R.; Baxter-Jones, C.; Bitgood, F.A.; Bradley, K.M.; *et al. Avian Genetic Resources at RISK. An Assessment and Proposal for Conservation of Genetic Stocks in the USA and Canada*; Report No. 20; University of California Division of Agriculture and Natural Resources, Genetic Resources Conservation Program: Davis, CA, USA, 1999. Available online: http://grcp.ucdavis.edu/publications/index.htm (accessed on 6 March 2012).

6. Delany, M.E. Genetic variants for chick biology research: From breeds to mutants. *Mech. Dev.* **2004**, *121*, 1169–1177.
7. Robb, E.A.; Delany, M.E. The expression of pre-axial polydactyly is influenced by modifying genetic elements and is not maintained by chromosomal inversion in an avian biomedical model. *Cytogenet. Genome Res.* **2012**, *136*, 50–68.
8. Gitter, C.L. Genetic Mapping of Seven Developmental Mutations Found in the Domestic Chicken (*Gallus gallus*) by Analysis of Single Nucleotide Polymorphisms. Master of Science thesis, University of California, Davis, Davis, CA, USA, December 2006.
9. Bellott, D.W.; Skaletsky, H.; Pyntikova, T.; Mardis, E.R.; Graves, T.; Kremitzki, C.; Brown, L.G.; Rozen, S.; Warren, W.C.; Wilson, R.K.; *et al*. Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. *Nature* **2010**, *466*, 612–616.
10. Darling, A.C.E.; Mau, B.; Blattner, F.R.; Perna, N.T. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **2004**, *14*, 1394–1403.
11. Robb, E.A.; Delany, M.E. Developmental Syndromes in the Chicken Biomedical Model: The Hunt for Causative Elements using Next-Generation Sequencing; Fine-Mapping; and Gene Expression Techniques. In *Proceedings of the 61st Annual American Society of Human Genetics (ASHG) and 12th International Congress of Human Genetics (ICHG) Joint Meeting*, Montreal; Canada, 11–15 October 2011.
12. Mardis, E.R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **2008**, *9*, 387–402.
13. Schuster, S.C. Next-generation sequencing transforms today's biology. *Nat. Methods* **2008**, *5*, 16–18.
14. Wheeler, D.A.; Srinivasan, M.; Egholm, M.; Shen, Y.; Chen, L.; McGuire, A.; He, W.; Chen, Y.J.; Makhijani, V.; Roth, G.T.; *et al*. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **2008**, *452*, 872–876.
15. D'Ascenzo, M.; Meacham, C.; Kitzman, J.; Middle, C.; Knight, J.; Winer, R.; Kukricar, M.; Richmond, T.; Albert, T.J.; Czechanski, A.; *et al*. Mutation discovery in the mouse using genetically guided array capture and re-sequencing. *Mamm. Genome* **2009**, *20*, 424–436.
16. Ng, S.B.; Turner, E.H.; Robertson, P.D.; Flygare, S.D.; Bigham, A.W.; Lee, C.; Shaffer, T.; Wong, M.; Bhattacharjee, A.; Eichler, E.E.; Bamshad, M.; Nickerson, D.A.; Shendure, J. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **2009**, *461*, 272–276.
17. Durbin, R.M.; Abecasis, G.R.; Altshuler, D.L.; Auton, A.; Brooks, L.D.; Gibbs, R.A.; Hurles, M.E.; McVean, G.A. A map of human genome variation from population-scale sequencing. *Nature* **2010**, *467*, 1061–1073.
18. Metzker, M.L. Sequencing technologies—The next generation. *Nat. Rev. Genet.* **2010**, *11*, 31–46.
19. Cosart, T.; Beja-Pereira, A.; Chen, S.; Ng, S.B.; Shendure, J.; Luikart, G. Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics* **2011**, *12*, doi:10.1186/1471-2164-12-347.

20. Kuchtey, J.; Olson, L.M.; Rinkoski, T.; MacKay, E.O.; Iverson, T.M.; Gelatt, K.N.; Haines, J.L.; Kuchtey, R.W. Mapping of the disease locus and identification of ADAMTS10 as a candidate gene in a canine model of primary open angle glaucoma. *PLoS Genet.* **2011**, *7*, doi:10.1371/journal.pone.1001306.
21. BGI Americas Homepage. Available online: http://bgiamericas.com/ (accessed on 13 March 2012).
22. Chakravarti, A. Genomic contributions to Mendelian disease. *Genome Res.* **2011**, *21*, 643–644.
23. Abplanalp, H. Inbred lines as genetic resources of chickens. *Poult. Sci. Rev.* **1992**, *4*, 29–39.
24. National Center for Biotechnology Information (NCBI) Homepage. Available online: http://www.ncbi.nlm.nih.gov/ (accessed on 13 March 2012).
25. National Center for Biotechnology Information (NCBI) *Gallus gallus* ftp (genome sequence) download site. Available online: ftp://ftp.ncbi.nih.gov/genomes/Gallus_gallus/ (accessed on 13 March 2012).
26. Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S.L. Ultrafast and memory efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **2009**, *10*, doi:10.1186/gb-2009-10-3-r25.
27. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079.
28. University of California Davis (UCD) Bioinformatics Core Homepage. Available online: http://bioinformatics.ucdavis.edu/ (accessed on 13 March 2012).
29. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760.
30. Burrows-Wheeler Aligner (BWA) Homepage. Available online: http://bio-bwa.sourceforge.net/ (accessed on 13 March 2012).
31. SAMtools Homepage. Available online: http://samtools.sourceforge.net/ (accessed on 13 March 2012).
32. Quinlan, A.R.; Hall, I.M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26*, 841–842.
33. BEDtools Homepage. Available online: http://code.google.com/p/bedtools/ (accessed on 13 March 2012).
34. Wang, J.; He, X.; Ruan, J.; Dai, M.; Chen, J.; Zhang, Y.; Hu, Y.; Ye, C.; Li, S.; Cong, L.; Fang, L.; Liu, B.; Li, S.; Wang, J.; Burt, D.W.; Wong, G.K.; Yu, J.; Yang, H.; Wang, J. ChickVD: A sequence variation database for the chicken genome. *Nucleic Acids Res.* **2005**, *33*, 438–441.