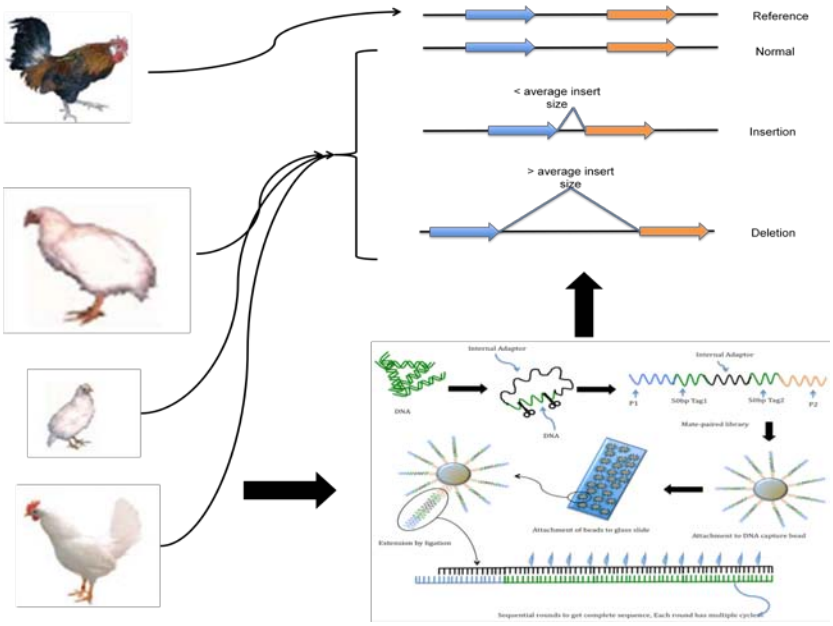
Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science

Bioinformatics analysis of whole genome resequencing data in the chicken

Khurram Maqbool



Department of Animal Breeding and Genetics
Examensarbete 333
Uppsala 2010

Master's Thesis, 30 HEC

One-Year Master's Programme in Biology – Bioinformatics

Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics

Bioinformatics analysis of whole genome resequencing data in the chicken

Khurram Maqbool

Supervisors:
Leif Andersson, SLU, Department of Animal Breeding and Genetics
Erik Bongcam-Rudloff, SLU, Department of Animal Breeding and Genetics
Carl-Johan Rubin, UU, Department of Medical Biochemistry and Microbiology
Examiner:
Göran Andersson, SLU, Department of Animal Breeding and Genetics

Credits: 30 HEC
Course title: Degree project in Animal Science
Course code: BI1021
Programme: One-Year Master's Programme in Biology
- Bioinformatics

Level: Advanced, A2E

Place of publication: Uppsala
Year of publication: 2010
Name of series: Examensarbete 333
Department of Animal Breeding and Genetics, SLU
On-line publication: http://epsilon.slu.se

Key words: Next generation sequencing; structural variation; insertion; deletion; chicken; whole genome resequencing; mate-pair

CONTENTS

ABSTRACT

One of the most important challenges in human medicine is the identification of genetic variants underlying complex diseases but we are still not able to assess the risk of disease by genetic profiling. Also, as the complete functional properties remain unknown for most genes it is very important to further functionally annotate genomes. The large abundance of domestic animals kept by humans constitute a valuable asset for the identification of genetic elements involved in complex disease and to unravel gene functions. Advantages of using animal models in studies of complex traits include the possibilities to strictly control environmental variation and to perform invasive measurements. Domestic animals have the additional advantage that different breeds segregate for traits of importance to human diseases.

Massive parallel sequencing technology (NextGen) enables researchers to resequence already completed genomes and enables whole-genome sequence information from multiple DNA samples to be determined in a short time. This makes it possible to, in a global manner, determine the extent of sequence polymorphisms among members within a species. Insertions and deletions are among the structural variation events that may be responsible for the variation in the expression of traits related to growth and reproduction in chickens. Rubin et al. (2010) pioneered the use of NextGen sequencing for population-based studies. The current study presented here will now extend previous analysis performed by Rubin et al., 2010 using paired reads, which makes identification of insertions/deletions in the chicken genome possible. Chicken lines whose sequences have been analyzed in this study include the red junglefowl as a reference bird, the High growth line (HL), the low growth line (LL) both established from White Plymouth Rock chickens in 1957 (Dunnington and Siegel, 1996) and the White Leghorn line L13 (WL_L13). Whole genome resequencing was performed on three lines of chickens where genomic DNA from 11 chicken individuals were pooled from each line to generate paired-end reads. Paired-end reads of 50 bp each were generated using AB-SOLiD^{TM} v3.0. The overall coverage of the reads was 20-25x. Most deletions were found for sizes ranging from 4kbp to 10kbp in the HL, LL and the WL_L13 and most insertions were detected for size ranging from 1.1kbp to 3.1kbp. The coverage from matching pipeline was performed on chromosome 13 that contains 326 regions detected as deletions using mate-pair reads. The identified deletions overlapping duplicated regions in the chicken genome were removed for chromosome 13 in HL and considering a median coverage equal to zero (of reads mapped by the matching pipeline) in putative deleted regions detected using paired reads resulted in 73 deleted regions. It is possible that most of the identified deletions are false positives and further refinement in the approach is required to increase the probability of finding only true deletions and insertions as well. This could be due to the artifacts in the library as well as the in the assembly of the sequenced chicken genome. Further analysis can be carried out to find the exact breakpoints of the deletions. We suggest the approach of taking the set of reads that are unmapped in the genome and align them again by splitting those into two parts and allowing a gap or an insertion between them, as reads spanning these features should be present in the genomes of populations bearing true deletions and insertions, respectively.

INTRODUCTION

Domestic chicken (*Gallus gallus domesticus*) as a genetic model

The domestic chicken constitutes an important protein source in the human diet in the form of meat and eggs. Many genetic/genomic studies of chicken have been performed with the aim of mapping the loci underlying differences in growth and fecundity traits (Abasht, et al. 2006). Domestication of chickens started around 6000BC in South East Asia. Red junglefowl (*Gallus gallus),* the wild Asian species and main contributor for domestication, has been subjected to selection for meat and egg production over a long period of time that diverged the bird into domestic chickens (*Gallus gallus domesticus*). The chicken has also been commonly included in studies of vertebrate evolution and in studies aimed at unraveling basic concepts of genetics and embryonic development. It is considered that a primitive reptile was the common ancestor of both, birds and mammals about 310 million years ago (Fig. 1). Hence, comparison of chickens with other vertebrates becomes crucial to understand the divergence of birds from mammals. The particular placement of chickens within taxonomy, their importance as a source of nutrition for human beings and development of chicken populations to understand principles of genetics, make chicken an important model organism for biological research (Khasmi, 2004; Hillier, 2004).

The chicken genome

In the year 2000, one of the major breakthroughs was the sequencing of the human genome (Lander, 2001). Other genome projects involved those sequencing the genomes of related species to obtain more knowledge about the origin of species. Therefore, whole genome sequences were determined not only to uncover different aspects of mechanisms of genetics in the genome including for phylogenetic studies as well as genomics and evolution of individual genes in human genome (Lander, 2001). Although it is quite informative to compare the mammalian genomes, the information obtained from birds' genomes would further elaborate current understanding about the origin of vertebrates and it will also help study the genomics of individual genes. Chicken genome was taken as the first representative of all bird species. Furthermore, it also became the first agricultural animal to have its complete genome sequenced (Hillier, 2004). The draft sequence of Red junglefowl (*Gallus gallus,* WUGSC 2.1/galGal3) contained around one billion base pairs with approximately 20,000-23,000 genes (Hillier et al., 2004). The diploid chicken genome has 78 chromosomes with 38 autosomes and two sex chromosomes. These chromosomes are classified as macro chromosomes and micro chromosomes depending upon their respective sizes. Unlike mammalian genomes, female of chickens are heterogametic having ZW sex chromosomes while male chicken carry ZZ sex chromosomes (Masabanda et al. 2004). Whole genome sequencing could assemble only 32 chromosomes while leaving some of the micro chromosomes that could not be included in the assembly (Hillier, 2004).

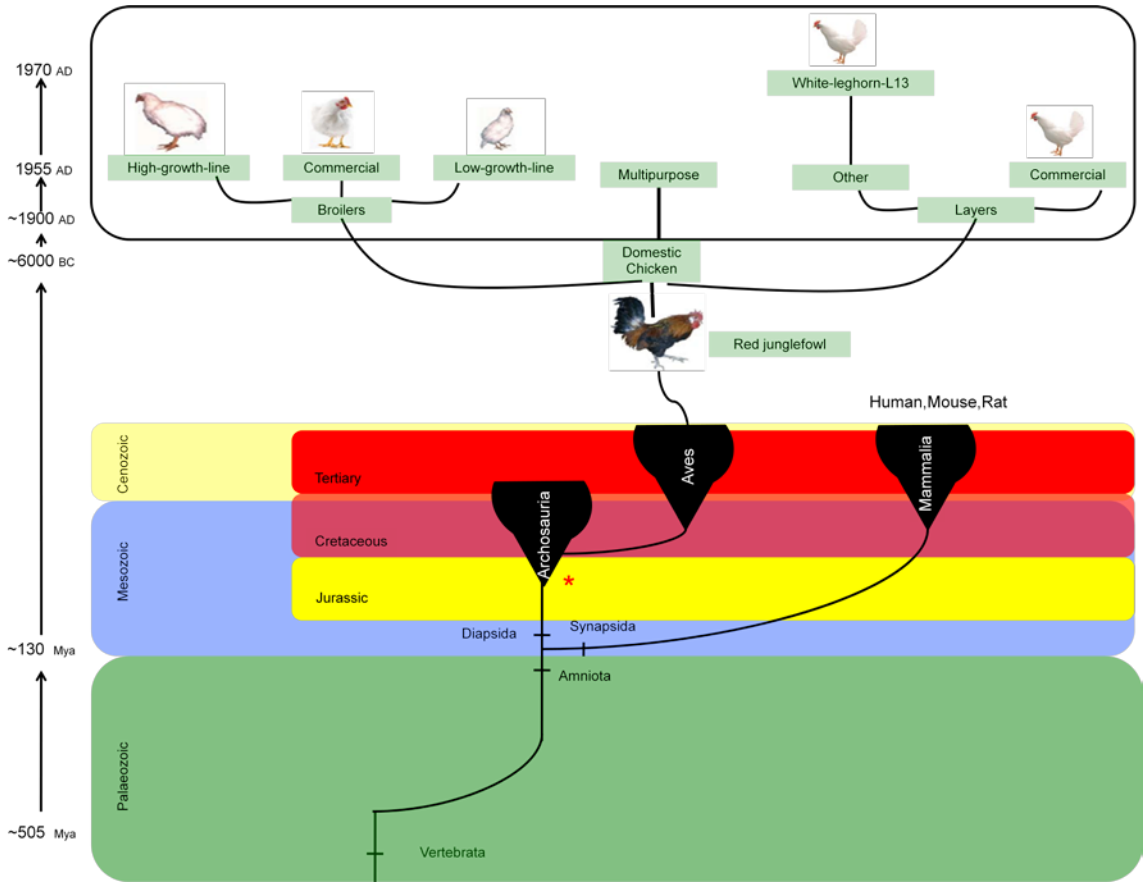


Figure 1. Bottom: Evolutionary relationship between birds (*Aves*) and mammals. The asterisk indicates the time when birds started diverge away from mammals after about 350Mya after the start of vertebrates (based on Hillier et al., 2004 and Rubin et al. 2010). The box at the top shows the result of the intense selection since the twentieth century, which generated two major branches of domestic chicken (broilers and layers). The populations whose sequences were analyzed in this study are included (*i.e.* the High growth line, the Low growth line and the White Leghorn Line 13.

Next generation sequencing

After the completion of the human genome sequence, the human genome marked major breakthrough in year 2001 (Venter et. al., 2001; Lander et al., 2001). Six years of characterization of sequence variation in the human genome began and this opened a whole new avenue to obtain more knowledge about the variation landscape of whole genomes (Pennisi, 2007). The Sanger sequencing approach was adapted to obtain the sequence draft of human genome and it took a relatively long time to sequence the whole genome because it involved bacterial artificial clones (BAC) of 100-kb DNA fragment that were used to amplify and generate approximately eight times coverage for each base pair over entire genome. Data was analyzed using information and knowledge from different sources to reconstruct human genome sequence "*in-silico*". The major advantage of this approach was that the quality of sequence was very high that gave just one error in approximately 40,000 bases in the human genome sequence. One of the major disadvantages of this technology is that it is both expensive and time consuming. Obtaining higher sequence coverage and an in-depth knowledge of the variations in the genome requires more samples to be sequenced and the

low throughput of Sanger sequencing makes it virtually impossible to sequence whole genomes for a large number of samples (Mardis, E. R. 2008; Shendure & Ji., 2008).

The development of Massive parallel sequencing (NextGen) made it possible to produce more sequence information from the comparative analysis of whole genome variants within a certain population than was possible with the previous approach of utilizing Sanger sequencing of BAC clones. It allowed for rapid and precise analysis of sequence polymorphisms as well as structural variations by scanning whole genomes. This technology is improving in terms of reducing the cost and time of whole genome sequence projects (Mardis, E. R. 2008). It also enables researchers to obtain multiple complete genome sequences for comparative analysis that can be used to identify the genes that are associated with different morphological, physiological and other important traits (Pennisi, 2007). Whole genome sequencing can be performed using different approaches. One of these is cyclic-array sequencing adapted by Roche (454 Genome sequencers, Roche Applied Science, Basal; http://www.454.com/enabling-technology/the-system.asp), Solexa/Illumina (Solexa/Illumina Genome Analyzer; Bently, 2006) AB-SOLiD™ (Applied Biosystems™), Helicos Heliscope™ (www.helicosbio.com) and Pacific Biosciences SMRT (www.pacificbiosciences.com). The SOLiD™ sequencing system from Applied Biosystems™ works in a similar fashion to Solexa/Illumina and Roche 454 Genome analyzer although there are some differences as it is based on sequence by ligation and each base on the read is called twice at each ligation cycle to introduce an inbuilt error control mechanism.

SOLiD mate-pair protocol

In the SOLiD™ mate-pair protocol, the library is constructed by developing random fragments of DNA of a specific size which are then selected to generate paired reads by applying an oligonucleotide adaptor based circularization (Fig. 2). This renders the mate-pairs joined by the internal adaptor and two adaptors named P1 and P2 are attached at the ends of the mate-pairs. The adaptor DNA complex is then attached to a 1 μm bead for amplification and then these beads are attached to individual positions on a glass slide. A universal primer of n bp anneals to one of the adaptors P1 or P2 and the sequencing is performed using DNA-ligase. A fluorescent-labeled eight base-pair fragment is then ligated to the primer where the fourth and fifth bases of the ligated fragment are fluorescent. The detector captures the fluorescence emitted after the cleavage of the eight bases ligated fragment. This step is repeated ten times to detect two bases at each cycle sharing the same color combination that is detected at each florescence emission. This is also referred to as two base encoding. This step could detect bases at positions 1-2, 6-7, 11-12, up to 46-47. A primer that ends at position n-1 at the 3' (the one position before the primer's 3' end attached in the last cycle) then replaces the previous sequence and the base positions shift to the left and again two bases are detected in ten cycles from each florescence. In the second step, the starting position becomes zero, one and it continues incrementing by four like 0-1, 5-6, 10-11, up to 45-46. After completion, the primer is replaced with n-2, n-3 and n-4 positions as mentioned above and for each of the primers, the ten cycles are repeated to obtain complete sequence of a read (Mardis, 2008; Shendure & Ji., 2008).

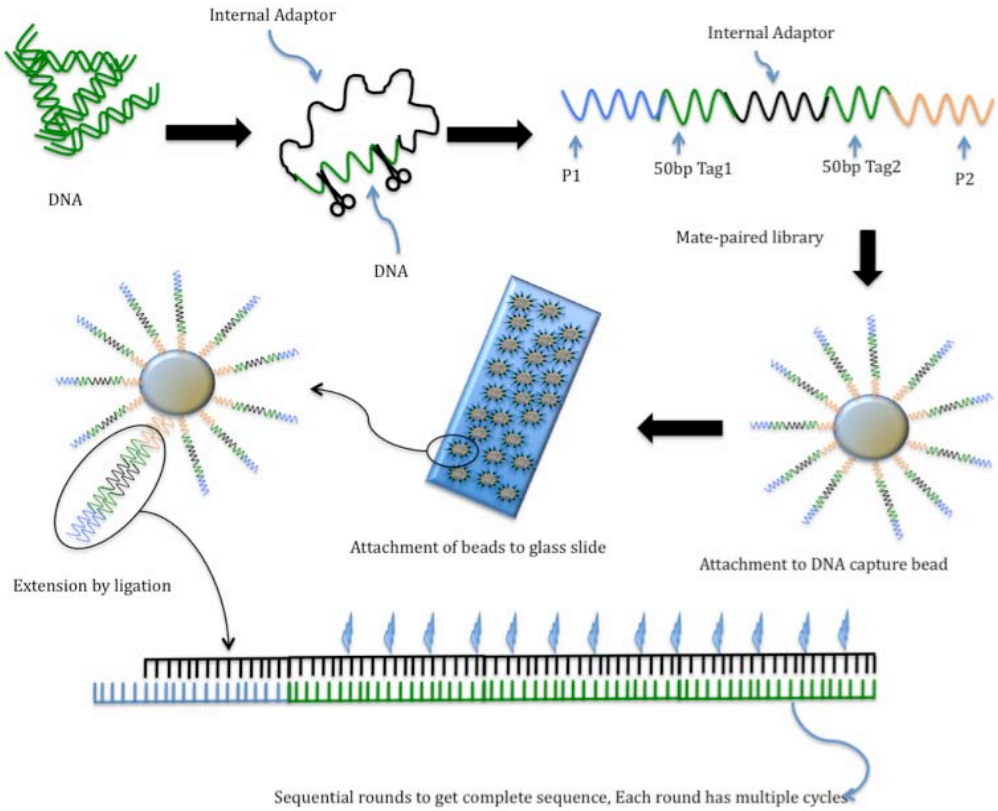


Figure 2. Overview of mate-pair sequencing using Applied Biosystems SOLiD™ technology. The DNA fragments are joined with an internal adaptor to generate mate-pairs. Two adaptors, P1 and P2, are attached at the ends of the mate-pairs and the complex is attached on 1 μm bead for amplification. The beads are attached on a glass slide where sequencing by ligation is performed and the two base encoding is then captured by the detector for each cycle of ligation in order to obtain a reading for each position of the reads (based on Mardis, 2008 and Shendure & Ji., 2008)

The raw data obtained from the instrument is sent to a computer for further analysis. In the analysis pipeline, paired reads of 50bp each are aligned to reference genome separately. The data mapped is combined to reconstruct mate-pairs giving orientation and distance between the mate-paired reads. Paired-reads that align at the expected distance and the expected orientation are called normal mate-pairs while some reads may deviate from the normal indicating the presence of structural variation in the sequenced genome compared to the reference genome (Applied Biosystems, 2008). Differences in the mate-pair distance of reads aligned to the reference genome from the normal mate-pair distance may indicate presence of insertions/deletions (indels) in the genome.

Whole genome resequencing of chickens

The use of NextGen sequencers will facilitate genetic studies of domestic animals enormously. Rubin et al. (2010) pioneered the use of NextGen sequencing for population-based studies. They sequenced pooled samples representing eight different populations of domestic chickens and zoo populations of the Red junglefowl making population-based studies affordable with current technology. The pooling approach in combination with next generation sequencing is a very cost-effective way to find regions of the genomes that show

striking differences in the allele frequency between populations. The study revealed 7.5 million single nucleotide polymorphisms (SNPs), more than 1,000 deletions occurring at a high frequency at least in one of the sequenced populations, and 38 loci with very strong signatures of selection. One selective sweep overlapped the *TSHR* gene encoding the thyroid-stimulating hormone receptor that has a key role for metabolic regulation as well as for regulating seasonal reproduction. A missense mutation in the coding sequence of *TSHR* was found in every tested domestic chicken representing more than 35 breeds of chicken with a geographic distribution from Iceland to China. Another major sweep signal was detected at the *TBC1D1* locus, encoding a protein involved in insulin signaling in muscle tissue. This sweep was specific for broilers (chickens used for meat production) and the gene overlaps the major QTL explaining differences in growth between broilers and layers (chickens selected for egg production) (Ambo et al., 2009). A third example is that a deletion of almost the entire *SH3RF2* gene was identified. The data strongly suggested that this is the causative mutation for a QTL affecting growth in a cross between High growth and Low growth chickens. *SH3RF2* is well conserved between birds and mammals but its function is completely unknown and this study establishes a new animal model to study its biological function. These examples illustrate how powerful the approach involving genome resequencing DNA pools is for detecting loci with clear signatures of selection.

Chicken was chosen for this first whole-genome sequencing study because it has a compact genome (about one Gbp), which made the study affordable already a year ago. However, since then the cost for sequencing has dropped dramatically as described above that makes it possible to resequence whole genomes. The reason for choosing domestic animals for comparative genome studies is that it would be beneficial to take advantage of the strong directional selection that has taken place in these species, which generates strong signatures of selection. Thus, the combination of whole genome resequencing data in domestic chickens combined with extensive genetic linkage- and gene expression data will provide unique opportunities to study genotype-phenotype relationships.

Indels among other structural variations play a key role in the divergence between closely related groups of organisms (Britten et al., 2003). They are among the structural variation events that can contribute to variation in traits related to growth and reproduction in chickens. Different methods can be applied to identify indels by analysis of genome resequencing data. If only single reads data are available then deletions can be identified on the basis of read coverage of different genome sequences (Rubin et al., 2010). Paired-end genome resequencing data gives more opportunities to scan the genome for identification of indels. Mate-pair libraries are made from fragments of known sizes. Reconstruction of the mate-pair distances *in-silico* can help identify indels using different approaches. Usually, short indels and large indels are identified separately because the algorithms for the two kinds of indels differ. Identification of short indels may be done be realigning the reads using less stringent alignment algorithms that tolerate gapped alignments . Large indels can be identified as regions where adjacent mate-pair reads are mapped to positions separated by a distance that significantly deviates from the normal insert size. Large insertions and deletions have been identified before in human genome using such an approach with mate-pair reads from SOLiD^{TM} system and posterior analysis performed by Corona Lite. In that study 1515 insertions and 4075 deletions were identified in human (McKernan et al., 2010).

The current study is aimed at extending previous analysis performed by Rubin et al., (2010), to generate a similar amount of additional data but instead of using single 35bp reads as were

used in the previous study, use 50bp mate-pair reads. These data will provide more information than the single reads for identification of variation among chicken lines, because it gives a much better power to identify structural rearrangements specifically insertions, deletions and inversions. Thus, the use of paired reads will also allow generating comprehensive lists of large insertion/deletions in the chicken genome and for instance search for frame-shift mutations that disrupt coding sequences.

AIM OF THE STUDY

A previous sequencing study used sequencing based on coverage from single reads to identify deletions in different populations of chickens (Rubin et al., 2010). The aim of the current study is to extend the analysis by using extensive bioinformatics analysis of paired reads and to be able to locate large insertions and deletions in three chicken populations; the HL, LL and the WL_L13. Identified putative insertions and deletions in the three chicken lines will be compared both with the published reference genome and with each other.

MATERIALS AND METHODS

Mate-pair library preparation

The mate-pair approach discussed in the previous section was used to prepare three independent libraries of pooled DNA from the HL, LL, and WL_L13 lines. Random fragments of the DNA were made and the size selection was performed where the inserts were derived from 3-4kb fragments. The middle part of the genomic DNA fragments has been digested away and 50bp from each end was available for sequencing.

Chicken resequencing data

SOLiD sequence were generated from pools of HL, LL and (WL_L13) chicken lines. Eleven males from WL_L13 were selected randomly at Swedish University of Agricultural Sciences (Liljidahl et al., 1979). Seven males and four females each from the HL and LL both established from White Plymouth Rock chickens in 1957 (Dunnington and Siegel, 1996) Sequencing was performed for each line after pooling genomic DNA from 11 individuals. Mate-pair reads of 50 bp each were generated using AB-SOLiD^{TM} v3.5. The matching pipeline (details of pipeline are given in the next section) gave 24.2Gbp of non-unique and unique mapped reads together and 20.5Gbp with only unique mapped reads. The overall coverage of the reads was thus 20-25x combined for the three populations.

Bioinformatic analysis

Corona Lite v4.2 (http://solidsoftwaretools.com/gf/) was used for aligning the reads to the chicken reference genome sequence (WUGSC 2.1/galGal3), which was generated from a single inbred female red junglefowl. The software provides a pipeline for the analysis of resequencing data obtained from the SOLiD^{TM} platform. The workflow of the analysis is given in Fig. 3. Primary analysis starts on the instrument that generates paired reads in the color-space fasta format called csfasta. The two mates of a read are designated as F3 and R3, each mate-pair is having its unique identification. After obtaining the paired reads in separate files both the mates of reads are aligned independent of each other against the reference genome. The base-space reference genome is converted to color-space fasta format for the alignment. The second step is to perform the pairing of the F3 and R3 reads using the Corona Lite pairing pipeline. The resulting files from the mapping pipelines serves as input for the pairing pipeline. Only unique matches are considered for constructing the mate-pair file. The

third step is to scan the mate-pairs for identification of putative insertions and deletions using Corona Lite large-indel-tool. The mate-pair library obtained from the pairing pipeline and the galGal3 reference genome are used as input for large-indel-tool pipeline. The Corona Lite-large-indel tool was tested and optimized for scanning insertions and deletions in the chicken genome on Isis and Grad clusters at the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX). The Isis and Grad clusters are running a queue management system called SGE (Sun Grid Engine). Using the procedure mentioned above, mate-pair libraries of 50 bp read length were obtained for three chicken lines (i.e. HL, LL and WL_L13) using the procedure mentioned above. The three chicken lines were scanned for putative indels. As a putative insertion or deletion we defined any region in the genome having clones from the paired library with insert sizes greater or shorter than average insert size and that would significantly deviate by a standard deviation of a value greater than six (Fig. 4).

S was calculated as follows, $S = \frac{d}{S.E}$

Where S, is the standard deviation, *d*, is the deviation in insert size from the average insert size for the whole genome and *S.E*, is the standard error for the whole genome based on standard deviation of all the normal reads and clone coverage.

Confidence intervals were calculated for the normal insert size considering 2.5 standard deviation to include the area from a normal distribution from the normal insert size for the three chicken lines separately. This was used to control the dispersion of the data group within the cluster of insertions and deletions by hierarchical clustering.

Corona Lite and Corona Lite Large-indel-tools were compiled and run on the “Isis” and “Grad” clusters at UPPMAX.

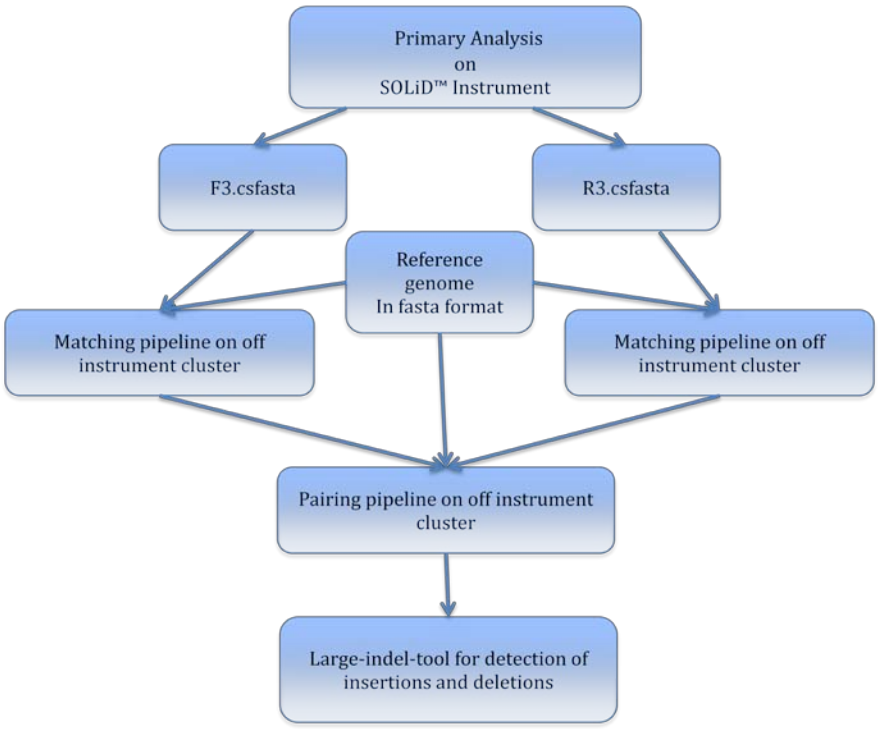


Figure 3. Corona Lite workflow for the chicken genome and the large-indel analysis pipeline using mate-pair reads obtained from SOLiD™ system

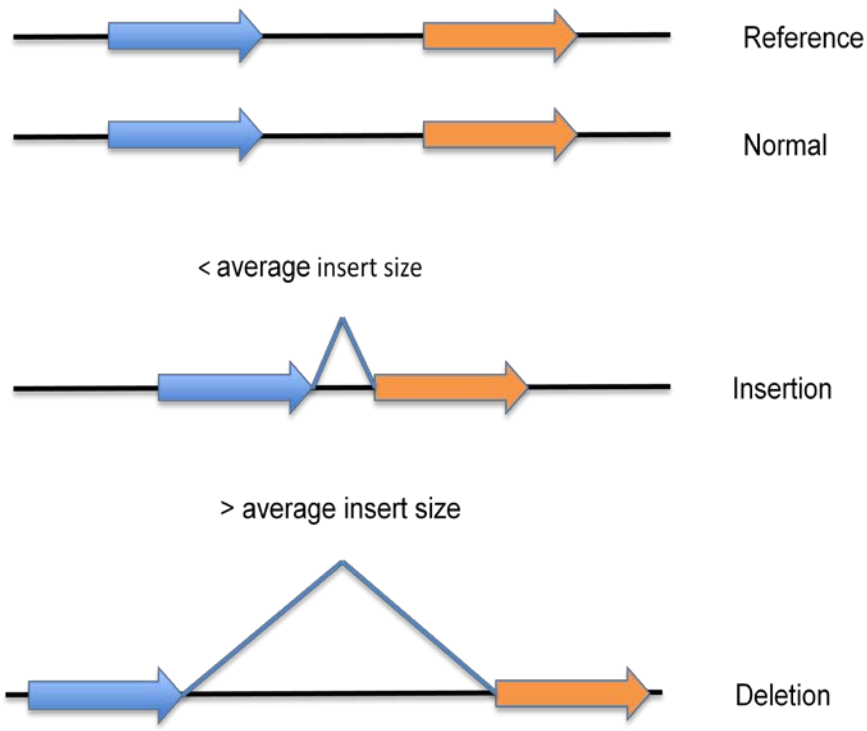


Figure 4. Characterization of insertion and deletion using mate-paired analysis. Blue and orange bars represent 50bp mate-paired sequence tags. The arrow represents direction of the sequence tags against the reference. Note that distances between mate-pair reads in the figure do not reflect true distances in the sequenced birds, but distances observed *in silico* in the mapping step in comparison with the reference genome against onto which reads were mapped.

Large-indel tool scripts

CORONALITE™ is the recommended software for the analysis of Applied Biosystems SOLiD™ sequencing color-space sequence. The main algorithms for analysis of the sequencing data are written in Perl (computer language) in the form of several scripts. After the installation of software, the main scripts used for analysis of sequencing data are stored in a folder named /bin in the installation folder of the software. The Large-indel-tool is integrated into the main pipeline of SOLiD™ software, Corona Lite. The scripts of the Large-indel-tool are also written in Perl and are used for scanning insertions and deletions in the genome sequencing data. The different steps of analysis are retained in different scripts. The analysis of genome sequencing data requires comparatively higher resources and needs cluster of computers making a super computer for proper execution of the analysis possible. The submission of scripts to queue for their execution is done using the different scripts for different operating system running on computer clusters. These scripts are "submit_scripts_to_SGE.pl", "submit_scripts_to_PBS.pl" and "submit_scripts_to_LSF.pl". Hence, the software only available to SGE (Sun Grid Engine), PBS (Portable Batch System) and LSF (Load Sharing Facility) managed clusters. At the time of this writing, there is no support available for one of the cluster management systems, SLURM (The Simple Linux Utility for Resource Management) within Corona Lite. The software runs by submitting the jobs at command prompt after defining the appropriate parameters for identification of insertions and deletions using mate-paired reads. The software then generates scripts using

those parameters and submits the scripts for execution to the queue manager. The current analysis was carried out on a computer cluster system UPPMAX. There are two queue management systems running at UPPMAX computer clusters, SGE and SLURM. The resources were used from UPPMAX that is running SGE operating system for scanning the chicken genome to find large insertions and deletions in the chicken genome. Because, Corona Lite does not support SLURM, all the analysis were tested, optimized and run only on SGE. Multiple scripts (as jobs) can be submitted to SGE by defining resources needed to execute the scripts. Initially Corona Lite v4.2 could not be successfully run on SGE because the job submission command line set in the software was not compatible with the standard command line accepted by SGE. Some modifications and optimizations in the scripts were needed to run the pipeline for chicken genome on SGE. Two changes were required to increase memory and time, one in script "submit_scripts_to_SGE.pl" and second in "large-indel-tool.pl". The first change is due to increase the memory resource in the script "submit_scripts_to_SGE.pl" at line number 154, as just increasing the memory size in the command line crashes the program. The second modification is for changing the command in the script "large-indel-tool.pl" at line number 289, to make it compatible with the SGE command line format at UPPMAX (Fig. 5). The term modified is "h_rt" represents the run time in the command line that represents time resource required to complete a job after the job starts running. The time should not be set to less than 30 hours to completely scan for insertions and deletions in the chicken genome including the largest chromosome in the genome.

The Corona Lite Large-indel-tool is designed to handle the human genome and hence it can analyze maximum 25 chromosomes including the 22 autosomes two sex chromosomes and one Un (unplaced) chromosome. Because the chicken genome sequence contains total of 32 chromosomes, the Large-indel-tool cannot perform a complete analysis of all the 32 chromosomes. Therefore, changes are required in the scripts of Corona Lite Large-indel-tool for the possibility to scan all the 32 chromosomes of chicken genome otherwise the analysis stops scanning reaching at 25^{th} chromosome. Also, the sex chromosomes are designated as W and Z in the chicken genome sequence data instead of X and Y in the human genome sequence data. The changes in the name of sex chromosomes also create a problem because the software needs exact numbers and names to be defined so that it can locate them properly. There are two scripts that contain the information for scanning the chromosomes in the genome, "preproc.pl" and "parse_and_add_chromosome.pl" both contained in the folder "~/CORONALITE/bin". Changing the respective values can make it possible to scan all the chicken genome having 32 chromosomes (Fig. 6). Modifications in the script "preproc.pl" are required at line numbers from 72 to 77 for number of chromosomes and name of the last three chromosomes. The last three chromosomes in the chicken genome are W, Z and N and take number 30, 31 and 32 respectively that are modified in the script. Here N is not an actual chromosome, but an artificial chromosome consisting of various unplaced sequences that were not placed on chromosomes in the chicken genome. Similar changes are required at lines from 42 to 75. In addition to modifications, some addition of the script is also required from line 42 to 56. Details of the modifications are given in figure 3 and figure 4. The updated version of the software is developed by name Bioscope^{TM}. This software may also require similar kind of changes for analysis of different genome sequencing data obtained from Applied Biosystems SOLiD^{TM} platform. Similar changes were made in Corona Lite in order to make it compatible with analysis of sequence data from different genomes.

submit_scripts_to_SGE.pl

```
154c154
<      return $opt_R if $opt_R =~ /4G/;
>      return $opt_R if $opt_R =~ /8G/;
```

large-indel-tool.pl

```
289c289
<      system("submit_scripts_to_SGE.pl -j $joblist");
>      system("submit_scripts_to_SGE.pl -P - - -    -R \"h_rss=8G,h_rt=40:00:0\" -j $joblist");
```

Figure 5. Changes for resources in two scripts: submit_scripts_to_SGE.pl & large-indel-tool.pl: blue lines specify the line numbers that are modified; red lines contain the original script and green lines correspond to mentions the modified script

preproc.pl

```
72,77c72,77
<       if ($chr1 eq '23') {
<           $ref = 'X';
<       } elsif ($chr1 eq '24') {
<           $ref = 'Y';
<       } elsif ($chr1 eq '25') {
<           $ref = 'M';
>       if ($chr1 eq '30') {
>           $ref = 'W';
>       } elsif ($chr1 eq '31') {
>           $ref = 'Z';
>       } elsif ($chr1 eq '32') {
>           $ref = 'N';
```

parse_and_add_chromosomes.pl

```
42,44c42,56
< for($i = 1; $i <= 25; $i++){
<     $file = $parameters->{assessment_directory} . "/" . $parameters->{mates_unique_file_name} . ".ref" . $i . "." . $parameters->{directory_extension};
<     open( FILE, "< $file" ) or die "Can't open $file : $!";
> for($i = 1; $i <= 32; $i++){
>    if ($i <=29 )
>   {$file = $parameters->{assessment_directory}."/".$parameters->{mates_unique_file_name}.".ref".$i."."$parameters->{directory_extension};}
>     elsif($i == 30)
>         {$file = $parameters->{assessment_directory} . "/" . $parameters->{mates_unique_file_name} . ".ref" . "W" . "." . $parameters->{directory_extension};}
>     elsif($i == 31)
>         {$file = $parameters->{assessment_directory} . "/" . $parameters->{mates_unique_file_name} . ".ref" . "Z" . "." . $parameters->{directory_extension};}
>     elsif($i == 32)
>         {$file = $parameters->{assessment_directory} . "/" . $parameters->{mates_unique_file_name} . ".ref" . "N" . "." . $parameters->{directory_extension};}
> open( FILE, "< $file" ) or die "Can not open $file : $!";
53c65
<       if($i >= 1 && $i <= 22){
>       if($i >= 1 && $i <= 32){
56,57c68,69
<       elsif($i == 23){
<           $assessments[$count]{chromosome} = 'chrX';
>       elsif($i == 30){
>           $assessments[$count]{chromosome} = 'chrW';
59,60c71,72
<       elsif($i == 24){
<           $assessments[$count]{chromosome} = 'chrY';
>       elsif($i == 31){
>           $assessments[$count]{chromosome} = 'chrZ';
62,63c74,75
<       elsif($i == 25){
<           $assessments[$count]{chromosome} = 'chrM';
>       elsif($i == 32){
>           $assessments[$count]{chromosome} = 'chrN';
```

Figure 6. Changes for chromosomes in two scripts: preproc.pl & parse_and_add_chromosomes.pl: blue line specify the line numbers that are modified; red lines corresponds to the original script and green lines contain the modified script

RESULTS

Large insertions and deletions in chickens

The mate-pairs both having positions in the same chromosome were scanned against the reference genome by analyzing the deviations in insert size across the genome in the HL, LL and WL_L13. A Confidence interval of 2.5 standard deviations away from the average insert size at both the tails was used to control dispersion of the data after calling the insertion and deletion. The HL and LL showed higher dispersion of normal reads as compared to WL_L13 (Fig. 7). Furthermore, the average insert size of WL_L13 was slightly lower than that of HL and LL. The confidence interval was calculated for each line separately (Table 1).

The confidence interval was also used to make clusters of insertions or deletions using hierarchical clustering, where the deviations of mate-pair distance from the average insert size was used to construct similarity index. This approach could identify high number of deletions as compared to insertions. Considering combined percentage of identified indels in HL and LL, 95% were deletions and 5% were insertions while the corresponding figures in WL_L13 were 88% deletions and 12% insertions. This could be inferred as WL_L13 having a lower fraction of deletions as compared to HL and LL. In analysis of the single read data previously generated, it was observed that the chicken reference genome contains a large number of falsely duplicated sequences (artificial duplications) (Zody et al. unpublished). Overlaps between herein identified indels and these artificial duplications revealed 8768, 8955 and 8502 in HH, LL and WL_L13 respectively total number of overlaps out of more than 26710 regions of artificial duplications for each of the three chicken lines analyzed (Table 1), corresponding to ~30 % of identified indels. Considering that the artificial duplications covered only 105 Mbp, an overlap of 30 % of indels is to a much higher fraction of overlaps than would be expected by chance. Thus, it is likely that the errors in the chicken genome assembly, such as the artificial duplications confer an important source of false positive indels identified in the current analysis.

Overlapping insertions between HL and LL were ~80% while only ~50% of insertions from WL_L13 overlapped with HL and LL. More than 80% of deletions overlapped among the three lines (Table 2). The insertion sizes ranged from 146bp to 5.64kbp in HL, 179bp to 3.54kb in LL and 141bp to 44.87kbp in White-Leghorn-L13 (Fig. 8). The size of deletions ranged from 2.57kbp to 99.13kbp in HL, 2.5kb to 100kb in LL and 1.92kbp to 97.03kbp in White-Leghorn-L13 (Fig. 9). Large fraction (~98%) of both insertions and deletions were homozygous (Table 1). Most deletions (~98%) were found in the 4kbp to 10kbp size range in the HL, LL and WL_L13 and most insertions (~96%) were detected in the 1.1kbp to 3.1kbp size range in HL, LL and WL_L13 (Fig. 8, 9). Large insertions of size greater than 4kb were more abundant in LL as compared to HL and WL_L13. Most paired-reads (~96%) deviating from normal insert size that gave support for evidence of insertion ranged from ten to 250 for insertions while most deletions (~98%) were supported by paired-reads ranging from ten to 300 in HL and LL while ~88% of the deletions were observed in the range from ten to 300. (Fig. 10,11). The standard deviation depends on insertion/deletion size and the numbers of paired-reads deviating from normal insert-size. Most standard deviations were in the range six to 20 for insertions and six to 45 for deletions (Fig. 12, 13). The deletions and insertions were identified using mate-paired reads, but the coverage of reads from matching pipeline on each base pair position in the region of deletions does not support all of the regions to be deleted as many had sequence reads mapped within the presumably deleted region. A sub-analysis was

performed for chromosome 13 in HL. The read coverage from matching pipeline was determined on chromosome 13 that contains 326 regions detected as deletions using mate-pair reads (Fig. 12). In a sub-analysis, the regions overlapping with the duplicated regions in the chicken genome were removed for chromosome 13 in the HL. Considering a median coverage equal to zero (of reads mapped by the matching pipeline) in putative deleted regions detected using paired reads resulted in 73 deleted regions (Fig. 15).

Identification of a true deletion

A large deletion over the *SH3RF2* gene was identified using the approach of absence of coverage in HL using single reads (Rubin et al., 2010). The current approach could also identify the deletions classified as homozygous from 18365846 bp to 18388549 bp on chromosome 13 (Fig. 16). The clones that deviated from a normal insert size that supported the presence of a deletion were 94 with this deletion having a high standard deviation (SD=255.5). The deviation of this deletion from the average insert size was 19147 bp.
In addition to this, smaller deletions in the same region were identified in LL and WL_L13 towards the extreme ends of the *SH3RF2* deletion region in HL. The exact positions for the first deletion in LL and WL_L13 were from 18367620 bp to 18371778 bp in LL and from 18367642 bp to 18371347 bp in WL_L13 with 91 and 107 number of deviated clones having a standard deviation of 11.23 and 16.74 for LL and WL_L13, respectively. The second deletion spans towards the other end of the *SH3RF2* deletion from 18379847 bp to 18385036 bp in LL and from 18379870 bp to 18384577 bp in WL_L13 with 91 and 69 number of deviated clones having standard deviation of 22.01 and 33.2 in LL and WL_L13, respectively.

Table 1. Statistic of insert size and frequency of large insertions & deletions in three chicken lines

	High-growth-line	Low-growth-line	White-leghorn-L13
Average insert size	3675	3625	3180
Standard deviation of insert size	726.6	740.1	459.7
Number of Insertions	1953	1806	3386
Number of Homozygous Insertions	1941	1790	3380
Number of Deletions	26167	26055	26757
Number of Homozygous Deletions	25553	25437	26287
Indels Overlapping with artificial duplications	8768	8955	8502
Total Indels	28121	27862	30144

Table 2. Comparison of insert sizes of overlapping and non-overlapping insertions and deletions among the three lines

	HL vs. LL	HL vs. WL	LL vs. HL	LL vs. WL	WL vs. HL	WL vs. LL
Not-Shared indels	3991	3836	3850	3880	5224	5400
Number of Insertions	487	332	355	265	1513	1580
Number of Homozygous Insertions	484	329	351	261	1513	1579
Number of Deletions	3504	3504	3494	3615	3710	3819
Number of Homozygous Deletions	3377	3337	3374	3430	3586	3707
Shared indels	24129	24284	24012	23981	24920	24744
Number of Insertions	1466	1621	1451	1541	1873	1806
Number of Homozygous Insertions	1457	1612	1439	1529	1867	1801
Number of Deletions	22663	22663	22561	22440	23047	22938
Number of Homozygous Deletions	22176	22216	22063	22007	22701	22580

Note: The data include redundant overlaps

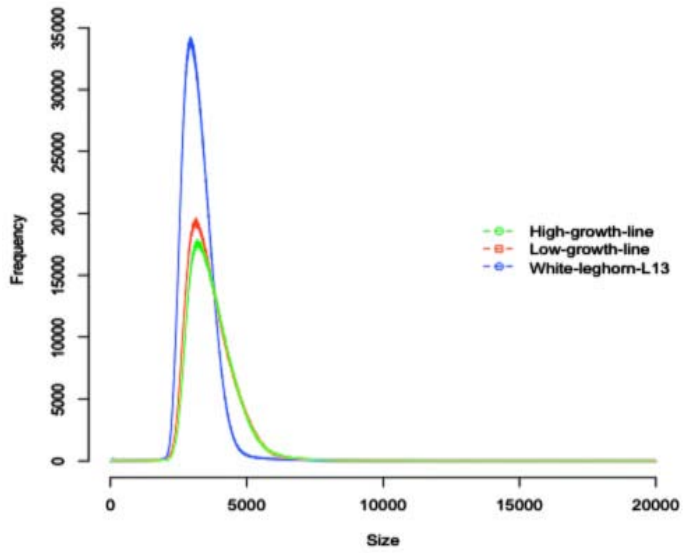


Figure 7. Histogram of insert sizes of the paired-reads in three chicken lines obtained from the pairing pipeline

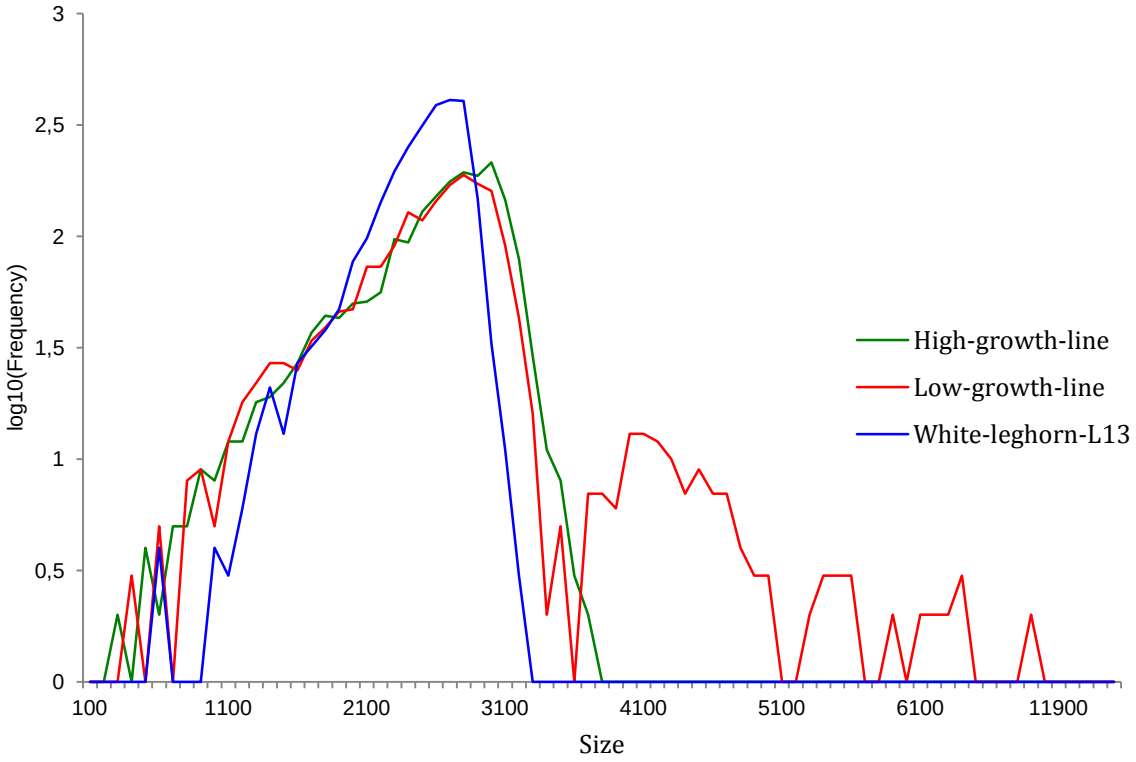


Figure 8. Histogram of insertion sizes identified using paired reads in three chicken lines. The x-axis contains discontinued range of the sizes of insertions to show only sizes of identified insertions.

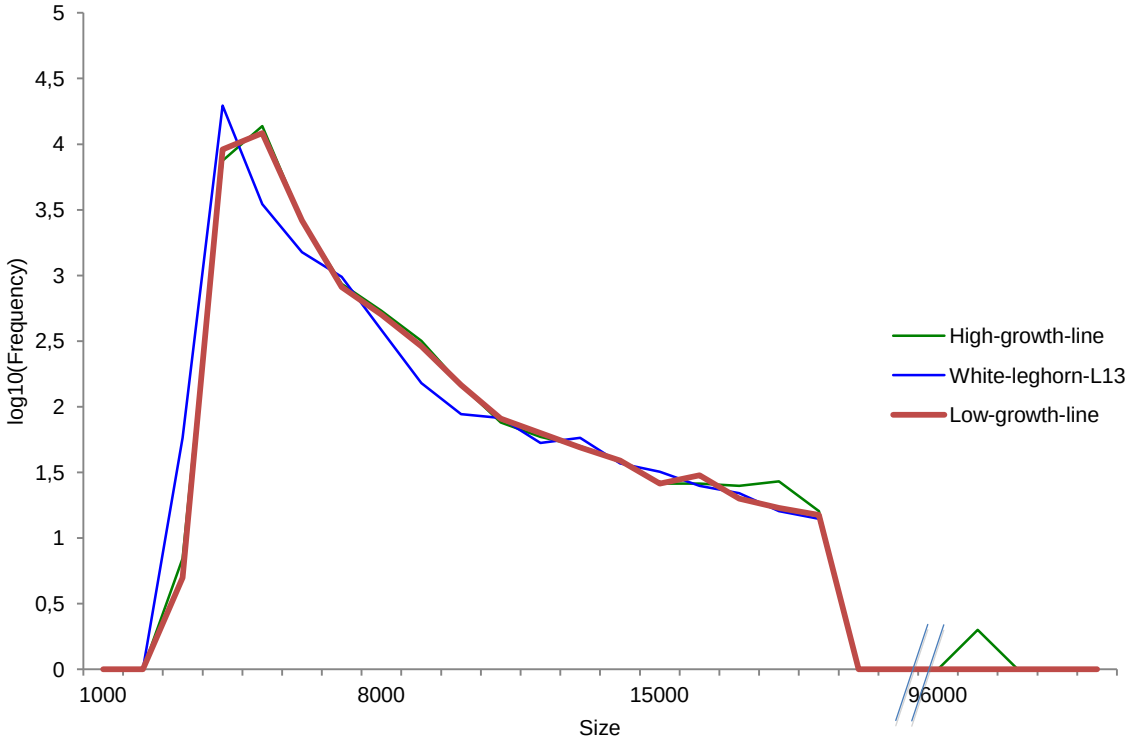


Figure 9. Histogram of deletion size identified using paired reads in three chicken lines The x-axis contains discontinued range of the sizes of deletions to show only sizes of identified deletions.

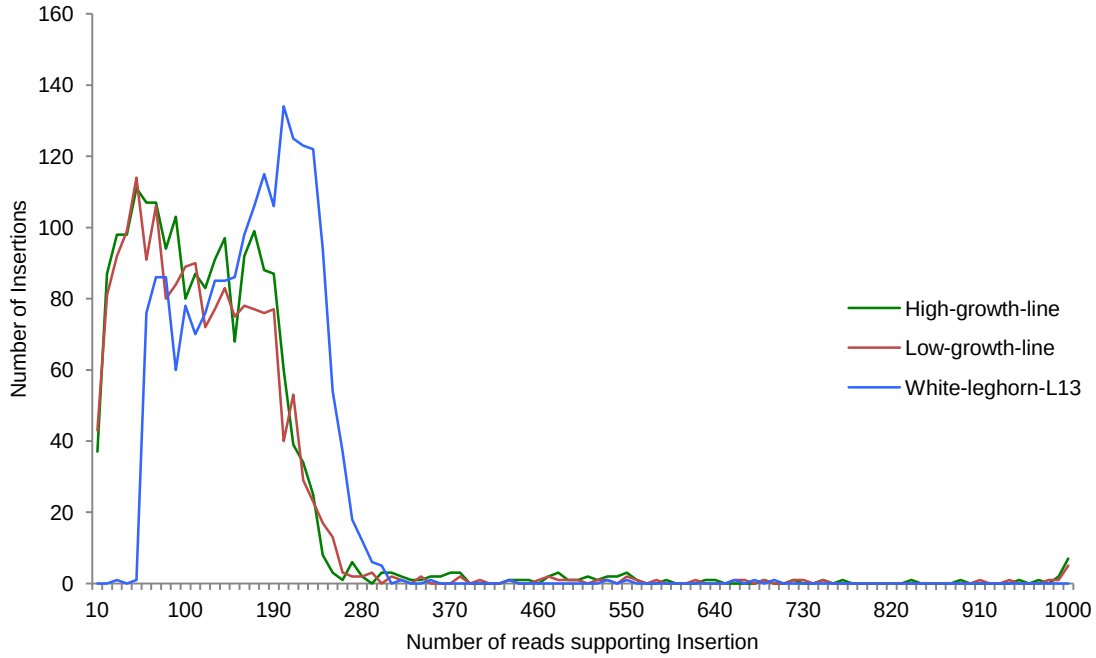


Figure 10. Histogram of mate-paired reads supporting insertions identified in the three chicken lines

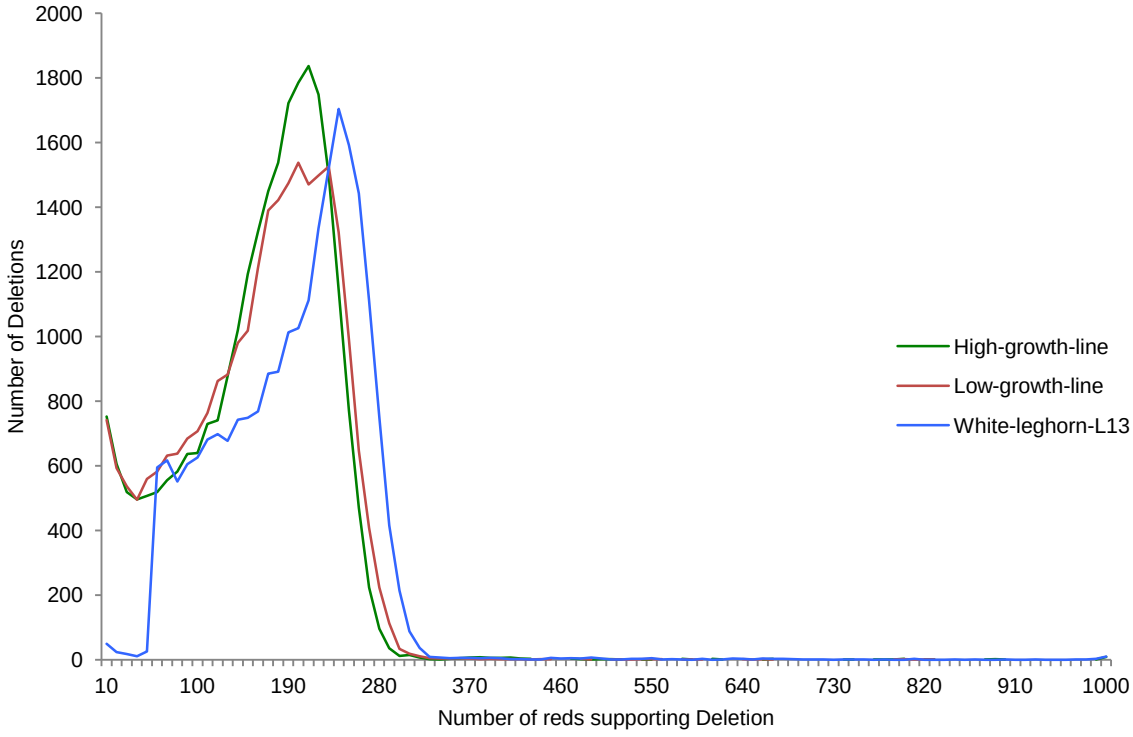


Figure 11. Histogram of mate-paired reads supporting deletions identified in the three chicken lines

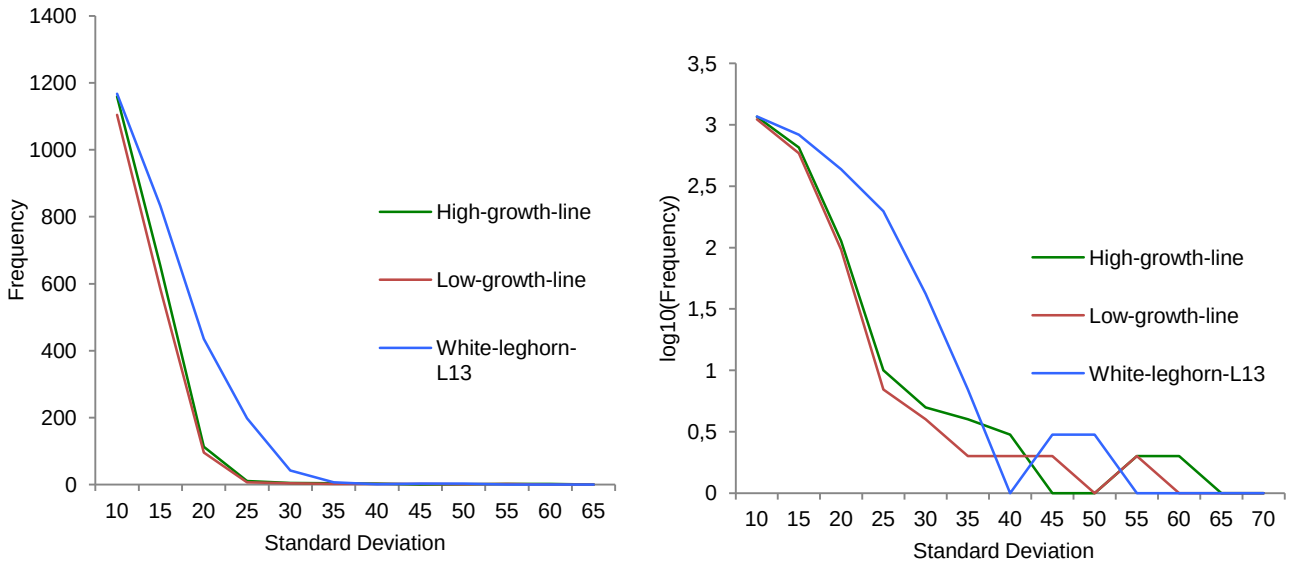


Figure 12. Left: Histogram of standard deviations, right: log10 of frequency histogram of standard deviations for insertion in three chicken lines

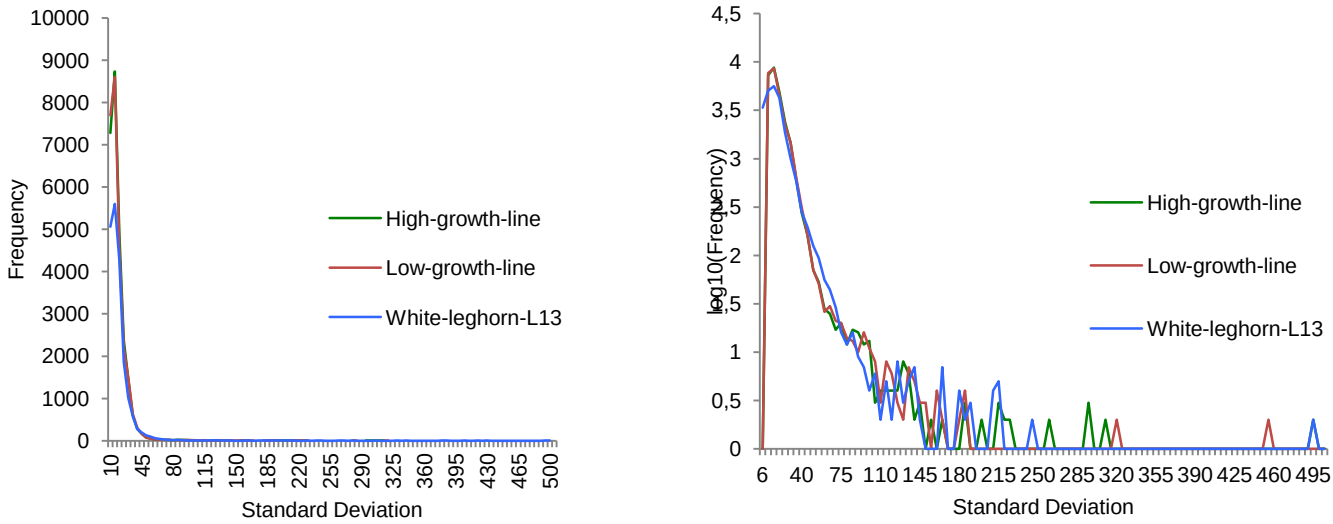


Figure 13. Left: Histogram of standard deviations, right: log10 of frequency histogram of standard deviations for deletions in three chicken lines

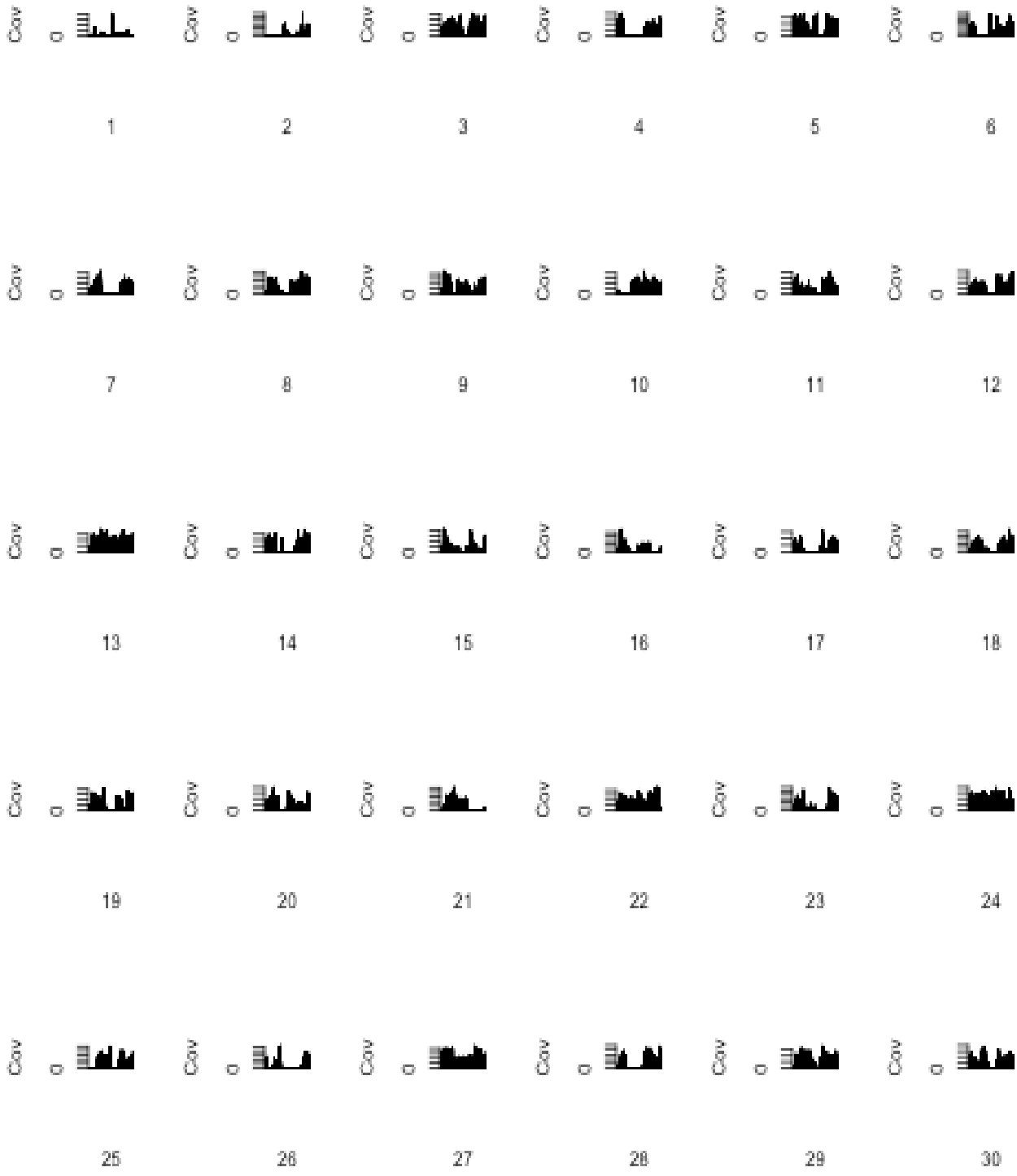


Figure 14. Coverage of reads on each base pair position from matching pipeline for a subset of the first 30 homozygous deletions on chromosome 13 from the total set of 326 identified in HL

Figure 15. Coverage of reads on each base pair position from matching pipeline for first 30 deletions selected on the basis of not occurring in duplicated region and median=0 on chromosome 13 from total set of 73 found in high-growth-line

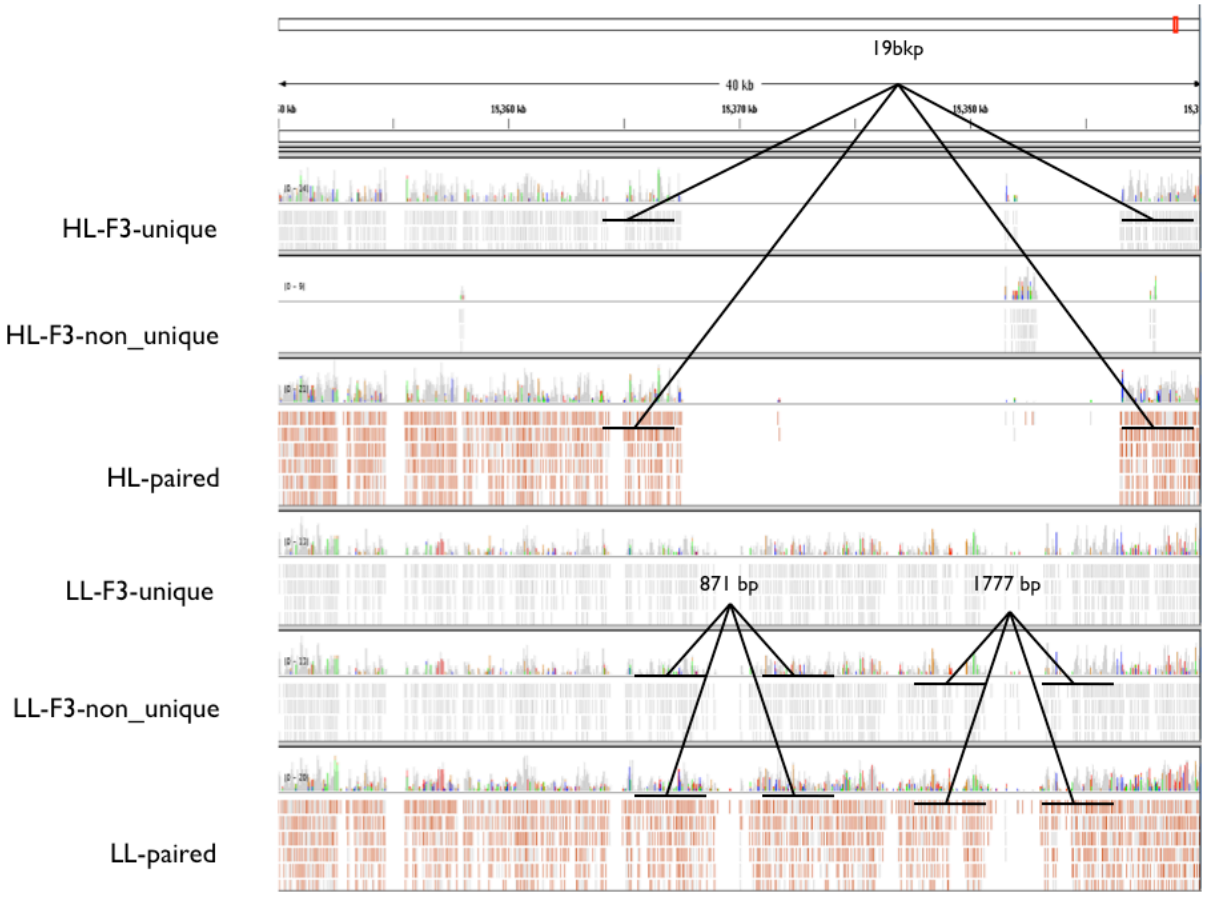


Figure 16. Coverage of reads on each base pair position from the matching (F3 unique and non-unique) and pairing pipeline (paired) for the SH3RF2 deletion region on chromosome 13 in the high-growth-line and the low growth line. Note larger deletion region in the high growth line (HL) for single reads as well s mate-pairs and two deletion regions in the the low growth line (LL) identified by Corona Lite large-indel-tool

DISCUSSION

Corona Lite is not suited to run on single processor and requires clusters of multi-processor nodes. It is also important to compile the software on the same machine that will be used to perform the analysis, as the compilation of the software is hardware dependent. The size of the chicken genome is 1.1Gbp (Hillier, 2004). The largest chromosome is chromosome 1 which is 200 Mbp. It takes a large amount of time to complete analysis by the pipeline and also requires sufficient memory resource for this chromosome. Therefore, memory and time resources for scanning the whole chicken genome for insertions and deletions need to be optimized in order to complete the analysis of larger chromosomes along with smaller chromosomes.

A large number of deletions were detected using approach as compared to the approach of considering the coverage of reads over the regions of genome to call deletions. Only 1300 deletions, larger than 100 base pair, were found in the domestic chickens (Rubin et al., 2010). Analysis of short insertions and deletions analysis revealed 140,484 polymorphisms in unique DNA (Brandström and Ellegren, 2007). It is possible that most of

the deletions are false positives and further refinement in the approach is required to increase the probability of finding preferentially true deletions and insertions. One of the possible confounding factors in the data set of indels detected using mapping distances between mate-pairs, could be the gaps and duplicated regions in the chicken genome, as such errors are not considered in the algorithm of large indel scanning. Most of the reads are discarded in the pairing pipeline because one of the pairs does not align in the genome. This could affect the scanning for insertions and deletions as most of the reads of normal insert size that are lost, will affect the coverage in the region of putative indel and hence the heterozygous indel could be identified as fixed putative indel. This can also lead to miscalculation of indel sizes. The use of pooled sample can lead to the identification of some of the indels as heterozygous because some of the individuals may be segregating for the non inserted/deleted allele. It is possible that some of the heterozygous indels may be identified as homozygous due to the use of pooled sample for the sequencing. The zygosity of an indel is also dependent on the confidence interval of the whole mate-pair library and the threshold of significance. This requires further analysis to optimize the parameters and some additional steps to identify true homozygous indels. In addition to the analysies using the large indel tool, further analyses were performed to check the coverage of reads from the matching pipeline in the regions identified as putative deletions by the Corona Lite large indel tool using mate-pairs from pairing pipeline. Some of them attained high coverage. Therefore, further analysis need to be carried out to exclude some of them if deemed to be false positive deletions. One of the possible reasons for detecting a high number of false positives is regions along duplications in the genome is because some of the mate-pairs match incorrectly in the pairing pipeline and are considered as significantly deviated from the normal insert size by the large-indel tool. Presence of large numbers of tandem duplications could also be detected in analysis of short insertions and deletions in chickens (Brandström and Ellegren. 2007).

There is also some additional filtering necessary of the reads on the basis of quality values (Sasson and Michael, 2010). Further analysis was performed to remove duplicated regions in chromosome 13, which eliminated 83 deletions while there was still coverage of reads from matching pipeline in the putative deletions identified by large-indel tool. Therefore, to eliminate those having coverage in the putative deletion region, the regions having a median coverage equal to zero were selected and these were found to be 73 on chromosome 13 in HL. Further analysis can be carried out to find the exact breakpoints of the deletions. We suggest the approach of taking the set of reads that are unmapped in the genome and align them again by splitting those into two parts and allowing a gap between them that is probably present in the genome assembly due to pertinent indels. More than 8000 indels found in the artificial duplicated regions in the three lines suggests the errors in the chicken genome assembly (Zody et al., manuscript). Furthermore, the high standard deviation of the insert size in the three chicken lines also limit the possibility of finding true indels.

FUTURE PROSPECTS

Most parameters in Corona Lite large-indel tool are optimized for the human genome by AB-SOLiD™; some scripts need changes in order to scan other genomes like the chicken for detection of insertions and deletions. Furthermore, the parameters in the Corona Lite large-indel tool defined to scan for insertion and deletion, particularly for confidence interval and standard deviation threshold also need to be recalculated for the chicken genome.

Indel detection using mate-pairs depends upon the reference genome assembly mapping, the algorithm used for aligning the reads to the reference genome and also the algorithm and the parameters to scan for the genome for identifying putative insertions and deletions. The matching algorithm and large-indel-tool of Corona Lite considers the chicken genome includes the gaps and the duplications in the genome while scanning for insertions and deletions. Artificial duplications as an artifact, has been identified in the chicken genome (Zody et al. unpublished). It is possible that the gap and duplication regions may compromise the identification of true indel. One of the possible approaches for identification of true indels is splitting the reads in equal parts of 23 bp each and realigning them with the reference genome to identify the breakpoints of the indels (Ameur ***et al.***, 2010). This may identify small number of reads because all the reads may not be equally separated after splitting from each other covering the deletion or insertion. The resequencing performed for each line gave approximately 7x coverage which is not sufficient to align significant number of reads that will identify breakpoints using split sequence approach. After identification of breakpoints several new artificial reference sequences, each corresponding to the sequence surrounding a putative indel breakpoint identified using the splitseq approach, could be constructed. These artificial references could then be used for mapping any unmapped reads that have a read pair which was mapped in the matching pipeline, using an alignment algorithm that tolerates insertions of gaps in the alignments. This will allow quantifying percentage of reads in the pool having deletion. The reads supporting the evidence of a deletion contribute to the statistical significance of the finding so this approach will also make it possible to identify homozygous deletions in the lines. Insertions and deletions reported in this study may have large number of false positives and alternative algorithms and approaches (considering the coverage from matching pipeline and split-sequence analysis) may be needed to find true significant putative insertions and deletions (Appendix).

ACKNOWLEDGEMENTS

I would like to acknowledge my supervisors Leif Andersson, Erik Bongcam-Rudloff, Carl-Johan Rubin for providing me constant guidance and an exciting research opportunity. Special thanks to Alvaro Martinez Barrio, Jonas Berglund, and Mohammad Kashif for helping me understand and resolve the programming problems. I would like to extent my thanks to my colleagues from Pakistan in the bioinformatics program. I am also grateful to Higher Education Commission, Pakistan for providing me the financial support to avail this opportunity. Special thanks to the resource and support from UPPMAX, the current study would not be possible without their support.

REFERENCES

Abasht, B., Dekkers, J.C. and Lamont, S.J. (2006) Review of quantitative trait loci identified in the chicken. Poult Sci. 85(12): 2079-96.

Ambo, M. et al. (2009) Quantitative trait loci for performance traits in a broiler 3 layer cross. Anim. Genet. 40, 200–208.

Ameur A, Wetterbom A, Feuk L, Gyllensten U. Global and unbiased detection of splice junctions from RNA-seq data. Genome Biol. 2010 Mar 17;11(3):R34.

Applied Biosystems (2008) SOLiD™ Systems mate-paired libraries detect and define large genetic rearrangements. Applied Biosystems: 139/AP06-01.

Bentley, D.R. (2006) Whole-genome resequencing. Curr. Opin. Genet. Dev,l 16:545–52.

Brandström, M. and Ellegren, H. (2007) The genomic landscape of short insertion and deletion polymorphisms in the chicken (*Gallus gallus*) genome: A High Frequency of deletions in tandem duplicates. Genetics, 176: 1691–1701.

Britten, R. J., Rowen, L., Williams, J. and Cameron, R. A. (2003) Majority of divergence between closely related DNA samples is due to indels. Proc. Natl. Acad. Sci. USA, 100: 4661–4665.

Crawford, R. D. (1990) Poultry Breeding and Genetics. Elsevier Science.

Dunnington, E. A. & Siegel, P. B. (1996) Long-term divergent selection for eight-week body weight in white Plymouth rock chickens. Poult. Sci. 75, 1168–1179.

Hillier, L.W., et al. (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature: 432, 695-716.

Khamsi, R. (2004) Chickens join the genome club: Battery of researchers hatch out full fowl sequence. Nature, doi: 10.1038/news041206-8.

Lander, E. S., Linton, L., M. Birren,. B. Nusbaum., Zody, M.C., et al. (2001) Initial sequencing and analysis of the human genome. Nature, 409:860–921

Liljedahl, L. E., Kolstad, N., Sørensen, P. & Maijala, K. (1979) Scandinavian selection and crossbreeding experiment with laying hens. I. Background and general outline. Acta Agric. Scand. 29, 273–286.

Mardis, E. R. (2008) Next-Generation DNA Sequencing Methods . Annu. Rev. Genomics Hum, 9: 387–402.

Masabanda, J.S., D.W.B., Patricia C. M. O'Brien, Vignal, A., Fillon, V., Walsh, P.S., Helen Cox, Tempest, H.G., Smith, J., Habermann, F., Schmid, M., Matsuda, Y., Ferguson-Smith, M.A., Crooijmans, R.P.M.A., Groenen, M.A.M., and Griffin, D.K. (2004) Molecular Cytogenetic Definition of the Chicken Genome: The First Complete Avian Karyotype. Genetics, 166: 1367-1373.

McKernan, K. J., Peckham, H. E., Costa, G., et al. (2010) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two base encoding. Genome Research, 19: 1527-1541.

Zody M. C. , Meadows, J. R. S., Sherwood, E., Sharpe, T., Rubin, C.J. , Eriksson, J., Lindblad-Toh, K. , Leif Andersson. Evaluation of the Chicken Genome Assembly with SOLiD Resequencing Data. (Manuscript)

Pennisi, E. (2007) Breakthrough of the year: Human genetic variation. Science: 318, 18.

Rubin, C., Zody, M.C., Eriksson, J., Meadows, J.R.S., Sherwood , E., Webster, M.T., Jiang, L., Ingman, M., Sharpe, T., Ka, S., Hallböök, F., Besnier, F., Carlborg, Ö., Bed'hom, B., Tixier-Boichard, M., Jensen, P., Siegel, P., Lindblad-Toh K., & Andersson L. (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. Nature, 464, 587–591.

Sasson, A. and Michael, P. M. (2010) Filtering error from SOLiD™ output. Bioinformatics, 26:849-850.

Shendure, J. & Ji, H. (2008) Next-generation DNA sequencing. Nature Biotechnology, 26: 1135-1145.

Venter J. C., Adams M. D., Myers E. W., Li P. W., Mural R. J., et al. (2001) The sequence of the human genome. Science, 291:1304–51

APPENDIX

Step-by-step protocol for identification of fixed insertions and deletions in chicken genome obtained from pooled data

The detailed flowchart of the analysis pipeline is given in the figure on next page

1. Make the mate-pair library from pooled samples of chicken population.
2. Run the matching pipeline using Corona Lite or Bioscope^{TM} for each F3 and R3 reads from mate-pair library. Use reference genome for alignment of the reads.
3. Combine and classify the reads from the matching pipeline to make mate-pairs in pairing pipeline. Use reference genome for mate-pair rescue.
4. Use the Corona Lite large-indel tool for the identification of large insertions and deletions. Use the reference genome for calculating the size of each chromosome.
5. Exclude the insertions and deletions falling in the duplicated and gapped regions of the reference genome.
6. Exclude the insertions and deletions having zero median value of clone coverage from the matching pipeline.
7. Run split sequence on all the reads to split the reads in parts and align them to the reference genome to identify the break points of insertions and deletions.
8. Combine the results obtained from Corona Lite large-indel tool, after applying the exclusion criteria, and split sequence analysis to identify true significant deletions with comparatively precise breakpoints.
9. Verify the insertions and deletions by using literature mining, and/or laboratory analysis.

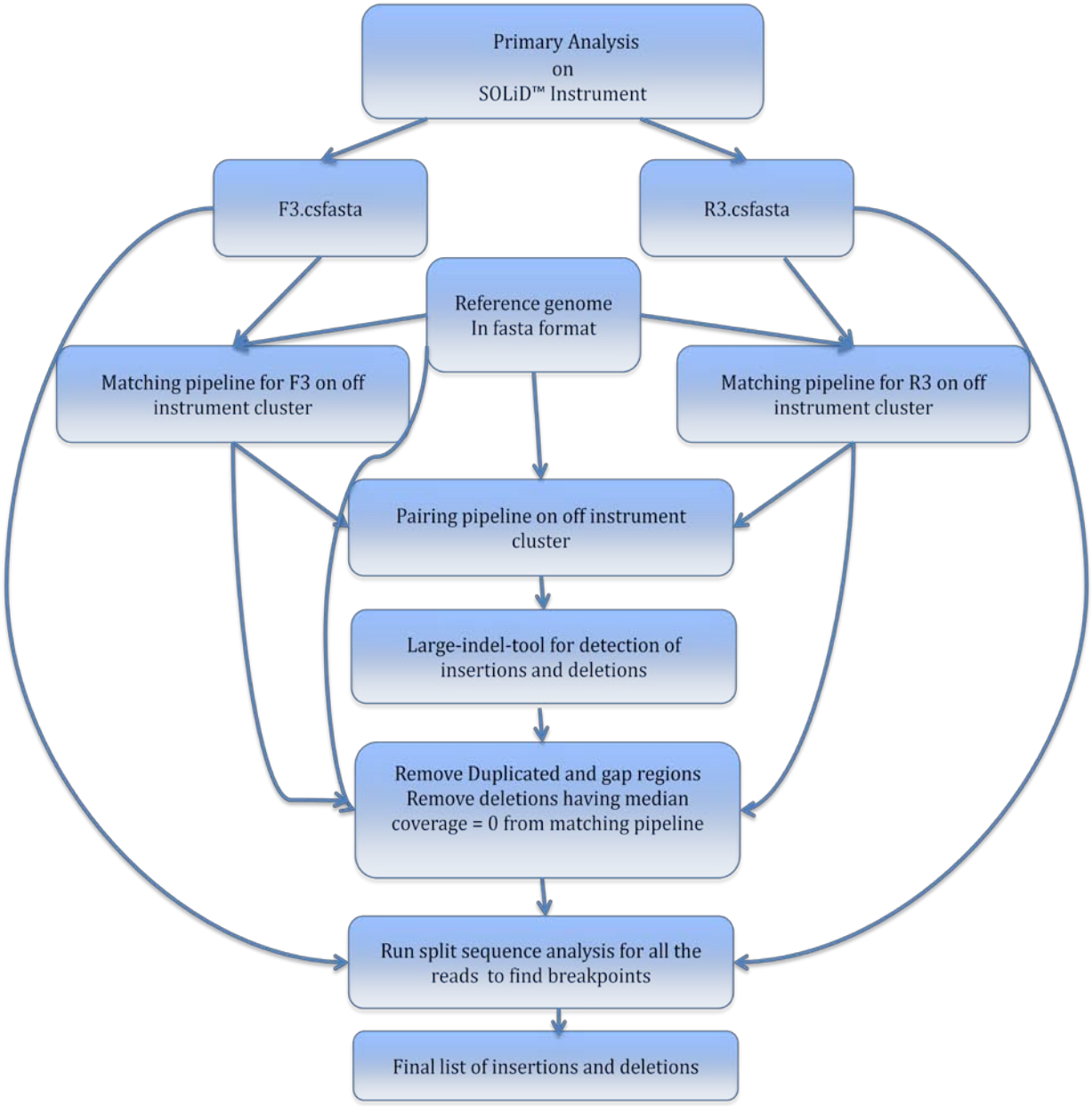


Figure 17. Proposed modified workflow for identification of large insertions and deletion analysis pipeline and remove false positives as well as identify breakpoints using Corona Lite large-indel-tool pipeline on mate-pair reads obtained with the SOLiD™ system for the chicken genome