BMC
Genomics

**RESEARCH ARTICLE**　　　　　　　　　　　　　　　　　　　**Open Access**

# Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage

Sandra Jansen[1], Bernhard Aigner[1], Hubert Pausch[1], Michal Wysocki[1], Sebastian Eck[2], Anna Benet-Pagès[2], Elisabeth Graf[2], Thomas Wieland[2], Tim M Strom[2,3], Thomas Meitinger[2,3] and Ruedi Fries[1*]

## Abstract

**Background:** Genome- and population-wide re-sequencing would allow for most efficient detection of causal trait variants. However, despite a strong decrease of costs for next-generation sequencing in the last few years, re-sequencing of large numbers of individuals is not yet affordable. We therefore resorted to re-sequencing of a limited number of bovine animals selected to explain a major proportion of the population's genomic variation, so called key animals, in order to provide a catalogue of functional variants and a substrate for population- and genome-wide imputation of variable sites.

**Results:** Forty-three animals accounting for about 69 percent of the genetic diversity of the Fleckvieh population, a cattle breed of Southern Germany and Austria, were sequenced with coverages ranging from 4.17 to 24.98 and averaging 7.46. After alignment to the reference genome (UMD3.1) and multi-sample variant calling, more than 17 million variant positions were identified, about 90 percent biallelic single nucleotide variants (SNVs) and 10 percent short insertions and deletions (InDels). The comparison with high-density chip data revealed a sensitivity of at least 92 percent and a specificity of 81 percent for sequencing based genotyping, and 97 percent and 93 percent when a imputation step was included. There are 91,733 variants in coding regions of 18,444 genes, 46 percent being non-synonymous exchanges, of which 575 variants are predicted to cause premature stop codons. Three variants are listed in the OMIA database as causal for specific phenotypes.

**Conclusions:** Low- to medium-coverage re-sequencing of individuals explaining a major fraction of a population's genomic variation allows for the efficient and reliable detection of most variants. Imputation strongly improves genotype quality of lowly covered samples and thus enables maximum density genotyping by sequencing. The functional annotation of variants provides the basis for exhaustive genotype imputation in the population, *e.g.*, for highest-resolution genome-wide association studies.

**Keywords:** Next-generation sequencing, Low-coverage, Genotyping by sequencing, Variant annotation

## Background

Association studies for the mapping of quantitative trait loci (QTL) and genetic evaluation in cattle is presently based on high-density SNP marker sets. The DNA variants that are causally involved in QTL variation are not necessarily among these markers. Since linkage disequilibrium of markers and causal variants is not complete, marker-based QTL mapping and evaluation are consequently not maximally efficient [1,2]. Maximal efficiency can only be achieved if all variable positions in the genome are assessed by whole-genome re-sequencing [3].

Although the advent of the next generation sequencing technology led to a dramatic reduction of sequencing costs, whole-genome sequencing of thousands of animals per population is still not affordable. Scheet and Stephens [4] proposed to sequence a subset of individuals and to impute the missing genotypes in the general population *via* genotypes obtained by a SNP chip. The

* Correspondence: ruedi.fries@tum.de
[1]Chair of Animal Breeding, Technische Universität München, Liesel-Beckmann-Strasse 1, 85354, Freising, Germany
Full list of author information is available at the end of the article

sequenced animals are ideally key ancestors of the population, *i.e.* unrelated individuals with a maximal contribution to the genomic variation of the modern population [5,6].

The feasibility and affordability of a whole-genome sequencing project also depends on the coverage required for efficient variant detection. A pilot of the Human 1000 Genomes Project, consisting of sequencing 179 individuals with a coverage of 3.6x per individual only, allowed to catalogue the vast majority of variation in this sample [7]. The low coverage prevented immediate and reliable variant calling at many sites. However, multi-sample analysis and imputation provided a means of highly accurate and complete genotyping, compensating for the low coverage [8,9].

The promising results of the first phase Human 1000 Genomes Project prompted us to apply low- to medium-coverage sequencing to key ancestors and some contemporary animals of a cattle population. Imputation within the sample of 43 sequenced animals and subsequent genotype imputation allowed the identification of more than 17 million variable sites, most of them novel, with high confidence.

## Results and discussion
### Selection of key animals for re-sequencing
Re-sequencing of a panel of optimally selected 43 key ancestors of the Fleckvieh population would enable to capture ~76 percent of the genetic diversity of the current population. However, obtaining DNA samples of key ancestors is often difficult. The animals died decades ago and semen samples are not always available. Thus, the actual 43 animals chosen for re-sequencing account for about 69 percent of the genetic diversity only.

However, the twelve most influential animals account for ~50 percent of the genetic diversity of the entire population (Figure 1). The set of selected animals comprises one sire-dam-offspring trio and totally eight parent-offspring duos (Additional files 1 and 2).

### Re-sequencing data
Close to ten billion paired-end reads of 36, 76, 96, 100 and 101 bases were obtained for the 43 animals on Illumina GA IIx and HiSeq2000 instruments. From 90.06 percent to 98.35 percent of the reads were mapped to the 30 chromosome scaffolds (autosomes 1 – 29, X-chromosome) of the bovine reference genome assembly UMD3.1 [10] encompassing 2.66 billion bases. Not considering PCR duplicates, that amounted from 3.59 to 15.55 percent (average 7.42 percent) of the reads, the coverage ranged from 4.19 to 24.98 reads per position, with an average of 7.46 reads for the 43 animals (Additional file 3). The re-sequencing data have been contributed to the 1000 bull genomes project [11].

### Detection of single nucleotide variants and short insertions/deletions
Multi-sample variant detection [12] allowed the calling of genotypes at 17,170,787 positions in each animal, meaning that at least one individual showed a deviation from the UMD3.1 reference sequence at each of these positions. In the case of missing or insufficient coverage, the most likely genotype was called based on the allele frequency at this site. *Beagle* [13] haplotype phasing and imputation raised the percentage of genotypes with phred-scaled quality scores greater than 10 from 94.92 (before imputation) to 98.34 (Figure 2). Most of the variants (89.91 percent) were biallelic single nucleotide
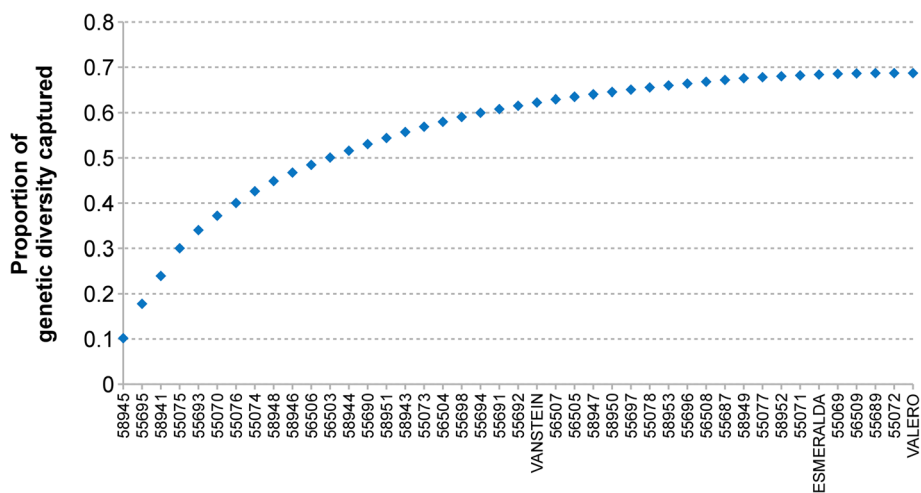


**Figure 1 Proportion of genetic diversity captured by re-sequencing of 43 key animals of the Fleckvieh population.** Forty-three animals chosen for re-sequencing are ordered based on their genetic contribution to the current Fleckvieh population. Among the re-sequenced animals is a trio including the previously re-sequenced *VANSTEIN* [36], his mating partner *ESMERALDA* and their common progeny *VALERO*.
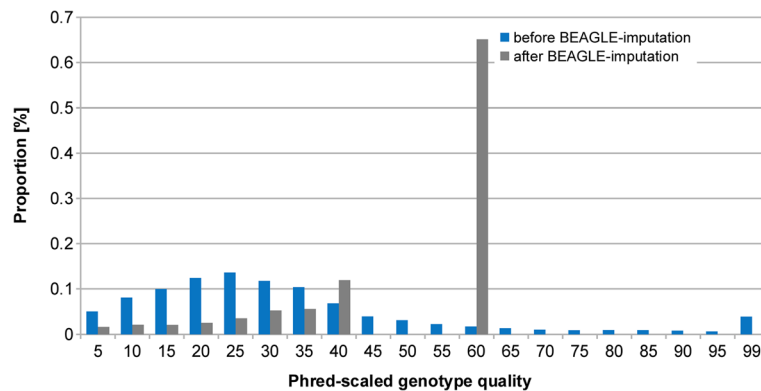
**Figure 2 Distribution of genotype-qualities before and after *Beagle* imputation.** The quality of each genotype was specified by its phred-scaled likelihood. The maximun score was restricted to 60 after *Beagle* imputation at 15,438,621 biallelic sites.

polymorphisms (SNPs) (Table 1). Insertion / deletion variants (InDels) amounted to 9.94 percent of the variants (54.45 percent deletions, 45.55 percent insertions). The remaining 0.15 percent were triallelic variants. Of the SNPs detected in this study, 67.95 percent represent newly identified variants; the remaining 32.05 percent are listed in build 133 of dbSNP. The rate of transitions to transversions (Ti/Tv) amounts to 2.21, which is within the range observed in previous whole-genome sequence studies [8]. The number of biallelic SNPs per animal ranged from 5,885,050 to 6,366,501 and did not depend on the coverage (Additional file 4). The proportion of biallelic SNP calls at sites without coverage ranged from 0.54 percent to 9.6 percent in the 43 animals (average: 3.47 percent).

Variants were submitted to dbSNP (National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD) under handle ID TZ_TUM (dbSNP Build ID: B138) and will be available from the NCBI SNP database website upon the next build in March 2013.

## Evaluation of variant-calling

We compared array-called genotypes with sequence-derived genotypes at corresponding positions and calculated the non-reference sensitivity (NRS) and the non-reference discrepancy (NRD) rates [8] to assess the sensitivity and specificity, respectively, of the variant calling. Both measures depend on the coverage, on whether or not *Beagle* imputation is applied and on the minimal phred-scaled likelihood requirement (Figure 3). The effects of *Beagle* imputation and minimal phred score requirement are most prominent if the coverage is below 7x. The NRD rate for the animal with the lowest coverage (4.17x) was 9.78 percent before and 5.17 percent after *Beagle* imputation. Highest NRS values at comparably low NRD values are achieved after *Beagle* imputation and after removal of genotypes with phred scores below 10.

## Functional annotation of variants

The annotated sequence, including 1,000 basepairs of putative promotor sequence, encompasses 18,444 genes and 55,753,651 basepairs of sequence and exhibits a total of 244,546 variants (Figure 4). The coding sequence is affected by 91,733 exchanges, 42,519 (46.33 percent) of which are non-synonymous. Kawahara-Miki et al. [14] reported a considerably higher rate (57 percent) for a Japanese animal. Our result is more in line with the findings by Stothard et al. [15] and Zhan et al. [16] of 42 percent of non-synonymous exchanges in an Angus and two Holstein animals. Fifty-eight percent of the non-synonymous substitutions are predicted to be "probably damaging" or "possibly damaging" (Additional file 5) by PolyPhen-2 and 575 (1.35 percent) are predicted to cause a premature stop codon, similar to reported findings [15,16]. Each of the putative stop codon introducing variants needs to be scrutinized by manual examination of the reads, the re-annotation of the affected genes and Sanger sequencing. *Bona fide* stop codon causing variants need to be typed in a representative sample. Absence of homozygosity may reveal undetected recessive disorders with nondescript or embryonically / fetally lethal phenotypes, with the latter manifesting themselves as reduced fertility [17,18].

The catalogue of annotated variants allows to examine whether any of the causal variants for bovine Mendelian traits / disorders listed in OMIA – Online Mendelian Inheritance in Animals [19] is segregating in the resequenced sample of Fleckvieh and therefore in the entire Fleckvieh population. The UMD3.1 reference coordinates could be determined for 59 of the OMIA entries (Additional files 6 and 7). The 2bp deletion in *MOCS1* responsible for Arachnomelia (OMIA 001541–9913; [20]) could be detected in a known heterozygous carrier of the defective variant. An amino acid exchange in *RASGRP2* has been proposed to be the causal variant for a bleeding disorder in a Simmental animal (Thrombopathia,

**Table 1 Counts of variants per chromosome**

| Chromosome | Biallelic | Triallelic | InDels |
|---|---|---|---|
| 1 | 953,116 | 1,594 | 106,808 |
| 2 | 765,523 | 1,540 | 86,271 |
| 3 | 689,653 | 1,061 | 74,819 |
| 4 | 727,730 | 1,059 | 81,188 |
| 5 | 711,291 | 1,236 | 78,518 |
| 6 | 715,956 | 1,306 | 81,170 |
| 7 | 647,618 | 944 | 72,861 |
| 8 | 662,730 | 929 | 73,740 |
| 9 | 605,302 | 884 | 69,531 |
| 10 | 628,109 | 968 | 68,543 |
| 11 | 594,105 | 1,102 | 64,693 |
| 12 | 658,305 | 1,117 | 70,418 |
| 13 | 467,872 | 977 | 50,653 |
| 14 | 496,824 | 852 | 54,644 |
| 15 | 563,600 | 736 | 60,307 |
| 16 | 479,831 | 785 | 52,312 |
| 17 | 454,188 | 722 | 51,040 |
| 18 | 378,391 | 604 | 40,948 |
| 19 | 370,300 | 787 | 39,877 |
| 20 | 449,443 | 744 | 49,207 |
| 21 | 410,810 | 612 | 44,478 |
| 22 | 335,787 | 520 | 36,763 |
| 23 | 420,876 | 676 | 42,587 |
| 24 | 371,270 | 632 | 40,919 |
| 25 | 269,464 | 430 | 27,099 |
| 26 | 316,133 | 507 | 34,077 |
| 27 | 287,872 | 470 | 32,574 |
| 28 | 290,731 | 482 | 31,154 |
| 29 | 350,281 | 665 | 36,219 |
| X (non PAR) | 285,956 | 486 | 44,949 |
| X (PAR) | 79,554 | 119 | 8,253 |

Multi-sample variant calling was performed as described in Methods. The X chromosome was divided into a pseudoautosomal region (PAR) and a non-pseudoautsomal region (PAR) (see Additional file 9).

OMIA 001003–9913; [21]). Eight of the re-sequenced animals carry the variant. It turns out that there are indeed animals in the Fleckvieh population being affected with the bleeding disorder. A genome-wide association study with six thrombopathic animals as case group and 43 re-sequenced animals as controls yielded a strong association signal on chromosome 29 (Figure 5). Subsequent autozygosity mapping in the affected animals revealed a common 6.3 Mb segment of extended homozygosity encompassing the location of *RASGRP2*

and corroborating recessive inheritance. Sanger sequencing confirmed that all thrombopathic animals are homozygous for the pertinent amino acid exchange (c.701T > C, p.L234P, Chr29:43599204). Among the affected animals are descendants of a re-sequenced bull. This bull did not appear to carry the variant. However, the relevant position is covered by two reads only, and after *Beagle* imputation, we could indeed observe heterozygosity that was subsequently confirmed by Sanger sequencing. This exemplifies both the power of re-sequencing key ancestors for the detection of genetic disorders and the importance of the imputing step at lowly covered sites. The Fleckvieh breed exhibits a recessively inherited red and white coat (relative to "dominant black") and is thus expected to be homozygous for the "red factor" causing deletion in the *MC1R* gene (OMIA 001199–9913; [22]). All but one animal are homozygous. The only heterozygous animal, carrying both the deletion and the wild type allele is red and white. The reference sequence, derived from red and white Hereford animals [23], also contains the wild type allele of the *MC1R* gene. A deletion variant of *PMEL* (OMIA 001545–9913) causing coat color dilution segregates in the re-sequenced animals at a frequency of 0.22 without apparent phenotypic effect. However, carriers of the variant allele are reported to produce progeny exhibiting a "diluted black" phenotype when crossed with black and white animals that carry the "dominant black" allele of *MC1R.* Thus, our findings are in line with Schmutz and Dreger's [24] discovery of interacting *MC1R* and *PMEL* alleles.

There are only few reports on the detection of causal variants for quantitative traits in cattle, so called QTNs [25]. A lysine-to-alanine substitution in acylCoA: diacyglycerol acyltransferase (*DGAT1*; c.694_695GC > AA, p.A232K, Chr14:1802265–1802266) has a major effect on the fat content of milk and milk yield [26,27]. The frequency of the milk-fat-enhancing and milk-yield-lowering lysine-allele is 0.02 in the re-sequenced animals. A variant in the growth hormone receptor gene (*GHR*), causing a phenylalanine-to-tyrosine substitution (c.836T > A, p.F279Y, Chr20:31909478) that affects milk yield and composition [28,29] is segregating in the re-sequenced animals, with the milk-yield-enhancing and content-lowering tyrosine-allele amounting to a frequency of 0.05. A missense mutation in the gene encoding ATP-binding cassette, sub-family G (WHITE), member 2 (*ABCG2*; c.1742A > C, p.Y581S, Chr6:38027010) affecting milk yield and composition [30] is not segregating in the re-sequenced Fleckvieh animals, *i.e.*, they are fixed for the milk yield-enhancing allele. Allelic variation of the prolactin receptor gene (*PRLR;* c.54_55GT > AC, p.S18N, Chr20:39115344–39115345) affects protein and fat yield [29]. The asparagine-allele is yield-enhancing and has a
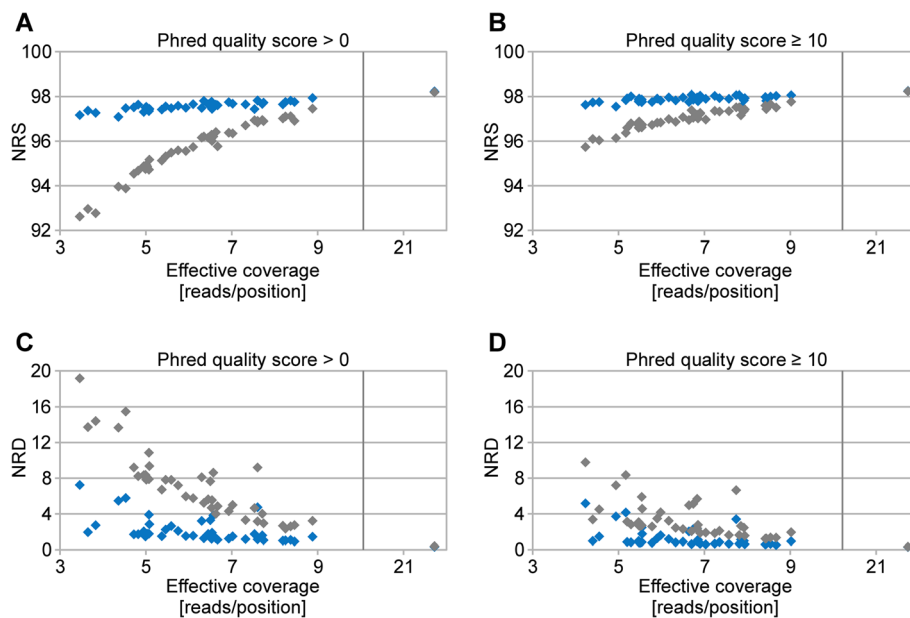
**Figure 3 Comparison of re-sequencing-derived with genotypes resulting from array-based genotyping.** Non-reference sensitivity (NRS), considering all genotypes (phred-scaled quality score ≥ 0) **(A)**. NRS, phred-scaled quality score ≥ 10 **(B)**. Non-reference discrepancy (NRD), phred-scaled quality score ≥ 0 **(C)**. NRD, phred-scaled quality score ≥ 10 **(D)**. The grey and the blue symbols represent the NRS and NRD before and after *Beagle* imputation, respectively.

frequency of 0.20 in the re-sequenced animals. Eberlein *et al.* [31] have shown association of a non-synonymous substitution in the G subunit of the non-SMC condensin complex (*NCAPG*, c.1326T > G, p.I442M, Chr6:38777311) with fetal growth. Our re-sequencing reveals the allele leading to reduced fetal growth at a frequency of 0.29. A recently postulated functional polymorphism close to the gene for pleiomorphic adenoma gene 1 (*PLAG1*) affecting stature [32,33] could not be assessed since it is located in a not-sequenced gap of the UMD3.1 reference sequence.

## Conclusions

Next-generation sequencing of key ancestors of a cattle breed and multi-sample calling as recommended by Nielsen [9] allowed the reliable detection of more than 17 million variable positions in the bovine genome, more than eleven million of them novel. Thus, this study significantly adds to the endeavour of sampling the genomic variation in *Bos taurus*. We show that genotype imputation within the re-sequenced animals enhances genotype quality considerably, particularly in lowly covered samples.
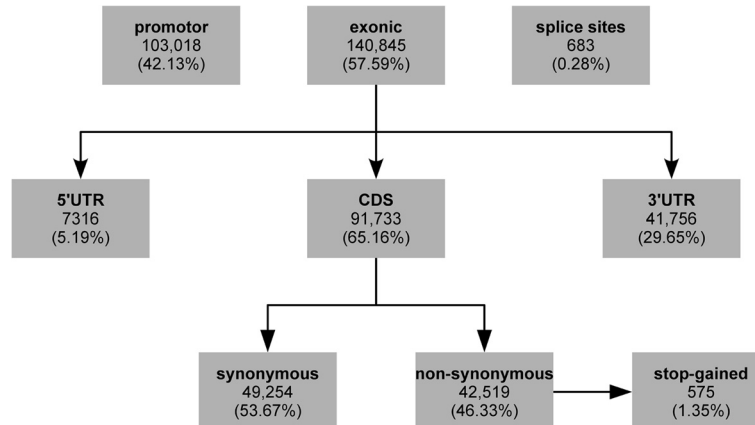


**Figure 4 Functional annotation of variants.** Functionally annotated biallelic SNPs and InDels in 18,444 genes (see Methods) including 1,000 bases of promotor sequence.
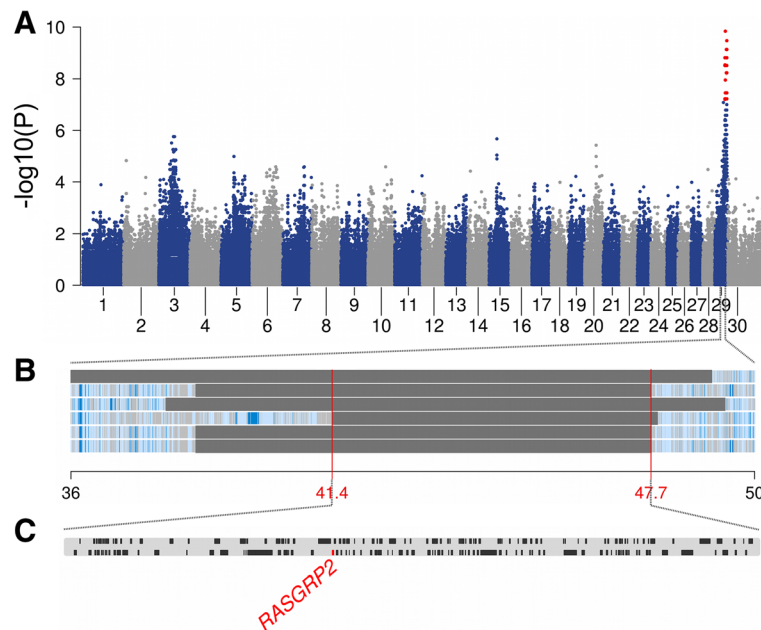
**Figure 5 Mapping of a bleeding disorder in the Fleckvieh population.** Association of 652,856 SNPs with the affection status of 43 re-sequenced and six thrombopathic animals **(A)**. The affected animals were reported to the Clinic for Ruminants, Faculty of Veterinary Science, University of Munich). P values were obtained using Fisher exact tests of allelic association. Autozygosity mapping in six thrombopathic animals revealed a common segment of extended homozygosity (41.3 – 47.7 Mb) **(B)** including the position of *RASGRP2* **(C)**.

The re-sequenced animals represent 69 percent of the genetic diversity of the Fleckvieh population. We can expect the majority of causal variants, at least the not very rare ones, to be segregating among the sequenced animals. If a causal variant can be pinpointed to a specific haplotype as a result of genome-wide association studies, and if the haplotype is present in the sequenced ancestors, identification will often be straightforward. Furthermore, inspecting already identified causal variants, such as listed in Online Mendelian Inheritance of Animals (OMIA, [19]), allows to readily monitor the population for genetic disorders. Since false positives are of no concern in this process, unfiltered variants should be considered and any match be confirmed by Sanger sequencing. Variant annotation led to the detection of 575 putative premature stop codons. Each of the underlying variants needs to be scrutinized with regard to possible annotation errors and to be confirmed by Sanger sequencing. Premature stop codons often represent loss of function mutations with potentially severe consequences such as embryonic lethality in the homozygous state. Thus, re-sequencing of key ancestors and functional annotation of variants should allow for a proactive management of genetic disorders.

The genotypes at more than 17 million positions can now be imputed in any animal of the Fleckvieh breed with array-derived medium or high density genotypes. Imputation accuracy should be high, at least for loci with minor allele frequencies > 0.05, since the re-sequenced animals, as key ancestors, are highly representative for the population [34]. However, increasing the number of re-sequenced animals will be necessary for maximal imputation accuracy for loci with lower minor allele frequency. One can now envisage to perform genome-wide association studies based on re-sequencing derived and imputed genotypes to achieve maximal resolution of QTL mapping and the perspective of direct identification of causal variants. The imputed genotypes will also be the basis for more accurately predicting genomic values of selection candidates [3].

## Methods

### Animal ethics statement

Semen samples were collected by approved commercial artificial insemination stations as part of their regular breeding and reproduction measures in cattle industry. The collection of blood samples was carried out by trained veterinarians during treatment of affected animals following standard veterinary protocols in Germany. No ethical approval was required for this study.

### Identification of key animals for sequencing

Genotype and pedigree data were available for 3645 animals representing an informative subset of the current German Fleckvieh population [35]. The genotyped animals were born between 1966 and 2009. Individuals

accounting for a major fraction of the genetic diversity of the available sub-population were considered as key animals. Key animals were identified iteratively with $p_m = A_m^{-1} * c_m$ [5], where $A$ is the numerator relationship matrix of $m$ selected animals with pedigree information lasting back to 1920, $c$ is a vector representing the average relationship of the selected animals to the entire population and $p$ is a vector with the proportion of genetic diversity captured by $m$ animals.

### Generation of sequence data

Genomic DNA was prepared from semen straws and blood samples following standard protocols using proteinase K digestion and phenol-chloroform extraction. DNA-concentration was set to 250 ng/µl. Paired-end libraries were prepared using the paired-end TruSeq DNA sample prep kit (Illumina inc., San Diego, CA, USA) and sequenced using Illumina GA IIx and HiSeq 2000 instruments (Illumina inc., San Diego, CA, USA). The read length ranged from 76 to 101 bp (Additional file 8). Paired-end reads with a length of 36 bp were available for one bull from a former study [36]. The resulting reads were processed with the Illumina BaseCaller during the sequencing step. The alignment of the reads to the University of Maryland reference sequence (UMD3.1) [10] was performed with BWA (version 0.6.1-r104) [37] using default parameters. The resulting per individual SAM files were converted into BAM files with SAMtools (version 0.1.18) [38]. Duplicate reads were identified and marked with the *MarkDuplicates* command of Picard [39].

### Variant-calling and imputation

Calling of variants including short insertions and deletions (InDels) was performed for the 43 sequenced animals simultaneously using *mpileup* of SAMtools (version 0.1.18) along with BCFtools [38]. The *extended base alignment quality* (BAQ) option was applied to scrutinize variant calling around InDels. The X chromosome was divided into a recombining pseudoautosomal region (PAR), and a non-pseudoautosomal region (non PAR) (Additional file 9). Read duplicates (see above) and positions with a coverage of more than 400 reads were not considered in variant calling. *Beagle* phasing and imputation [13] within the 43 sequenced animals was used to improve the primary genotype calling by SAMtools. The resulting variants were compared with those listed in the latest version of dbSNP (build 133; ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/cow_9913/chr_rpts/).
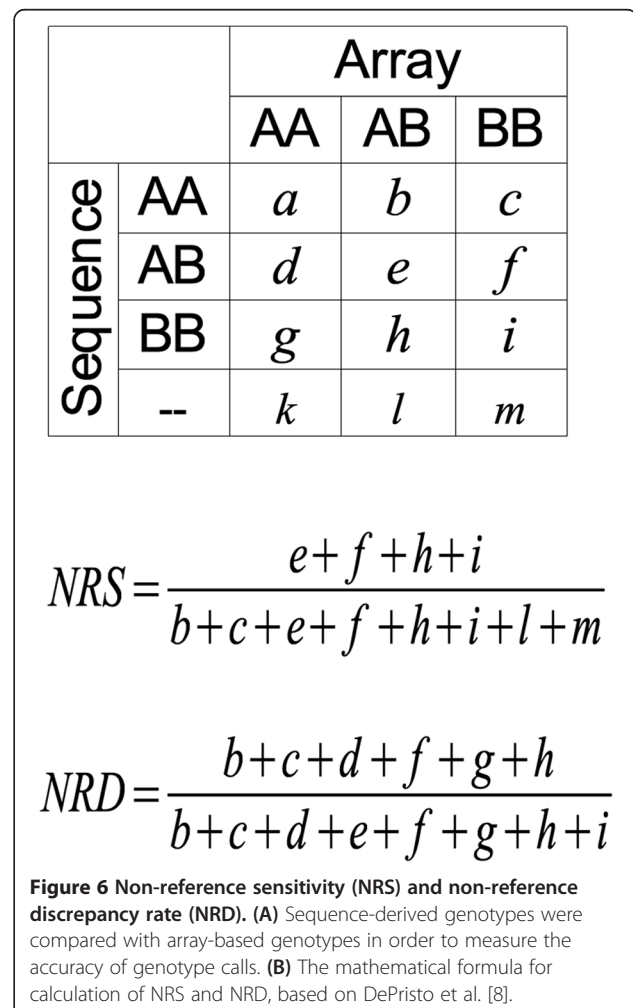
### Evaluation of variant-calling

All 43 re-sequenced animals were genotyped with the Illumina BovineHD BeadChip®. The average per individual call rate was 99.26 percent. The array-derived

genotypes were compared with the sequence-derived genotypes at in average 46,268 positions on chromosome 1 which we considered as representative for the autosomal genome. Non-reference sensitivity (NRS) and non-reference discrepancy (NRD) rates were calculated as proposed by DePristo *et al.* (2011) [8] and as shown in Figure 6.

### Variant annotation

The functional effects of unfiltered variants were analysed based on the annotation of the UMD3.1 assembly of the bovine genome with 21,364 protein coding transcripts [40]. Gene models leading to one or more in-frame stop codon were not considered, restricting the variant annotation to 18,444 protein coding transcripts (one per gene), including 32 manually annotated or re-annotated transcripts (Additional file 10). Positioning and analysis of variants in the promotor (defined to encompass 1,000 bp upstream of the transcription start), in the 5'-UTR, in the amino-acid coding region, in splice sites and the 3'-UTR were performed using a database of unfiltered variants set



$$NRS = \frac{e+f+h+i}{b+c+e+f+h+i+l+m}$$

$$NRD = \frac{b+c+d+f+g+h}{b+c+d+e+f+g+h+i}$$

**Figure 6 Non-reference sensitivity (NRS) and non-reference discrepancy rate (NRD). (A)** Sequence-derived genotypes were compared with array-based genotypes in order to measure the accuracy of genotype calls. **(B)** The mathematical formula for calculation of NRS and NRD, based on DePristo et al. [8].

up in MySQL and by custom scripts written in Python. *PolyPhen2-2 v.2.2.2* [41] was applied for predicting effects of non-synonymous mutations.

## Availability of supporting data

The identified variants were submitted to the Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. (dbSNP Build ID: {138}). Available from: http://www.ncbi.nlm.nih.gov/SNP/

## Additional files

**Additional file 1: Pairwise relationship of the 43 animals.** The pairwise relationship of the 43 re-sequenced animals was extracted from the numerator relationship matrix. Different color indicates the extent of relationship.

**Additional file 2: Ancestry of the re-sequenced animals.** The animals are ordered according to their birth year. Bold type indicates animals with re-sequenced sires.

**Additional file 3: Mapping statistics.**

**Additional file 4: Counts of biallelic SNPs per animal.**

**Additional file 5: Prediction of functional effects.** The prediction of effects of non-synonymous variants (n=42,519) was performed by using *Polyphen2-2 v.2.2.2* [41].

**Additional file 6: UMD3.1 coordinates of OMIA bovine causal variants.**

**Additional file 7: References of articles cited in Additional file 6.**

**Additional file 8: Distribution of read lengths per sequenced animal.** For each animal, one library was constructed. Eight animals were sequenced in two or more different runs resulting in different read lengths.

**Additional file 9: Determination of the boundary between the pseudoautosomal and non-pseudoautosomal regions on the bovine X chromosome.** High-density genotypes (777,962 SNPs per animal) were available for 896 male animals of the Fleckvieh population. The chromosomal position of the SNPs was determined according to the University of Maryland assembly of the bovine genome (UMD3.1). Animals and SNPs with call rates < 90 percent were omitted. The proportion of heterozygous animals was calculated for 38,868 SNPs on the X chromosome. The shaded box highlights the pseudoautosomal region that was estimated to extend from base 137,000,000 to the distal chromosome end because of the increased heterozygosity in this stretch. The pseudoautosomal region was subsequently treated like an autosomal area.

**Additional file 10: Newly annotated genes.** For 32 genes the genomic structure was predicted based on the University of Maryland UMD3.1 assembly of the bovine genome sequences [10] and the Dana-Farber Cancer Institute bovine gene index release 12.0 [42] by using GENOMETHREADER software tool [43]. The GENOMETHREADER output was viewed and edited using the Apollo sequence annotation editor [44]. Gene: gene symbol; Chr (strand): chromosome (orientation of gene); txStart: position transcription start; txEnd: position transcription end; cdsStart: position coding start; cdsEnd: position of coding end; exons: number of exons.

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

RF, TMS and TM conceived the study and participated in its design and coordination. MW, ABP and EG performed the sample and library preparation and carried out the sequencing experiments. SJ, BA, HP, SE, TW, TMS and RF

analysed the data. SJ and RF drafted the manuscript. All authors read and approved the final manuscript.

## Author details

[1]Chair of Animal Breeding, Technische Universität München, Liesel-Beckmann-Strasse 1, 85354, Freising, Germany. [2]Institute of Human Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764, Neuherberg, Germany. [3]Institute of Human Genetics, Technische Universität München, 81675, Munich, Germany.

## References

1. Meuwissen TH, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157**:1819–1829.
2. Goddard ME, Hayes BJ: **Mapping genes for complex traits in domestic animals and their use in breeding programmes.** *Nat Rev Genet* 2009, **10**:381–391.
3. Meuwissen T, Goddard M: **Accurate prediction of genetic values for complex traits by whole-genome resequencing.** *Genetics* 2010, **185**:623–631.
4. Scheet P, Stephens M: **A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase.** *Am J Hum Genet* 2006, **78**:629–644.
5. Goddard ME, Hayes BJ: **Genomic selection based on dense genotypes inferred from sparse genotypes.** *Proc Assoc Advmt Anim Breed Genet* 2009, **18**:26–29.
6. Gusev A, Shah MJ, Kenny EE, Ramachandran A, Lowe JK, Salit J, Lee CC, Levandowsky EC, Weaver TN, Doan QC, Peckham HE, McLaughlin SF, Lyons MR, Sheth VN, Stoffel M, De La Vega FM, Friedman JM, Breslow JL, Pe'er I: **Low-pass Genomewide Sequencing and Variant Inference Using Identity-by-descent in an Isolated Human Population.** *Genetics* 2012, **190**:679–689.
7. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467**:1061–1073.
8. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ: **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nat Genet* 2011, **43**:491–498.
9. Nielsen R, Paul JS, Albrechtsen A, Song YS: **Genotype and SNP calling from next-generation sequencing data.** *Nat Rev Genet* 2011, **12**:443–451.
10. Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS, Marçais G, Roberts M, Subramanian P, Yorke JA, Salzberg SL: **A whole-genome assembly of the domestic cow, Bos taurus.** *Genome Biol* 2009, **10**:R42.
11. *1000 Bull Genomes Project.* [http://www.1000bullgenomes.com/].
12. Li H: **A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.** *Bioinformatics* 2011, **27**:2987–2993.
13. Browning BL, Browning SR: **A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals.** *Am J Hum Genet* 2009, **84**:210–223.
14. Kawahara-Miki R, Tsuda K, Shiwa Y, Arai-Kichise Y, Matsumoto T, Kanesaki Y, Oda S, Ebihara S, Yajima S, Yoshikawa H, Kono T: **Whole-genome resequencing shows numerous genes with nonsynonymous SNPs in the Japanese native cattle Kuchinoshima-Ushi.** *BMC Genomics* 2011, **12**:103.

15. Stothard P, Choi J-W, Basu U, Sumner-Thomson JM, Meng Y, Liao X, Moore SS: **Whole genome resequencing of Black Angus and Holstein cattle for SNP and CNV discovery.** *BMC Genomics* 2011, **12**:559.

16. Zhan B, Fadista J, Thomsen B, Hedegaard J, Panitz F, Bendixen C: **Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping.** *BMC Genomics* 2011, **12**:557.

17. Vanraden PM, Olson KM, Null DJ, Hutchison JL: **Harmful recessive effects on fertility detected by absence of homozygous haplotypes.** *J Dairy Sci* 2011, **94**:6153–6161.

18. Charlier C, Agerholm JS, Coppieters W, Karlskov-Mortensen P, Li W, de Jong G, Fasquelle C, Karim L, Cirera S, Cambisano N, Ahariz N, Mullaart E, Georges M, Fredholm M: **A deletion in the bovine FANCI gene compromises fertility by causing fetal death and brachyspina.** *PLoS One* 2012, **7**:e43085.

19. *Online Mendelian Inheritance in Animals, OMIA.* University of Sydney: Faculty of Veterinary Science; 2012. World Wide Web URL: Http://omia.angis.org.au/.

20. Buitkamp J, Semmer J, Götz K-U: **Arachnomelia syndrome in Simmental cattle is caused by a homozygous 2-bp deletion in the molybdenum cofactor synthesis step 1 gene (MOCS1).** *BMC Genet* 2011, **12**:11.

21. Boudreaux MK, Schmutz SM, French PS: **Calcium diacylglycerol guanine nucleotide exchange factor I (CalDAG-GEFI) gene mutations in a thrombopathic Simmental calf.** *Vet Pathol* 2007, **44**:932–935.

22. Klungland H, Våge DI, Gomez-Raya L, Adalsteinsson S, Lien S: **The role of melanocyte-stimulating hormone (MSH) receptor in bovine coat color determination.** *Mamm Genome* 1995, **6**:636–639.

23. Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, Weinstock GM, Adelson DL, Eichler EE, Elnitski L, Guigó R, Hamernik DL, Kappes SM, Lewin HA, Lynn DJ, Nicholas FW, Reymond A, Rijnkels M, Skow LC, Zdobnov EM, Schook L, Womack J, Alioto T, Antonarakis SE, Astashyn A, Chapple CE, Chen H-C, Chrast J, Câmara F, Ermolaeva O, Henrichsen CN, *et al*: **The genome sequence of taurine cattle: a window to ruminant biology and evolution.** *Science* 2009, **324**:522–528.

24. Schmutz SM, Dreger DL: **Interaction of MC1R and PMEL alleles on solid coat colors in Highland cattle.** *Anim Genet* 2013, **44**(1):9–13.

25. Mackay TF: **Quantitative trait loci in Drosophila.** *Nat Rev Genet* 2001, **2**:11–20.

26. Winter A, Krämer W, Werner FAO, Kollers S, Kata S, Durstewitz G, Buitkamp J, Womack JE, Thaller G, Fries R: **Association of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-CoA:diacylglycerol acyltransferase (DGAT1) with variation at a quantitative trait locus for milk fat content.** *Proc Natl Acad Sci USA* 2002, **99**:9300–9305.

27. Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P, Cambisano N, Mni M, Reid S, Simon P, Spelman R, Georges M, Snell R: **Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition.** *Genome Res* 2002, **12**:222–231.

28. Blott S, Kim J-J, Moisio S, Schmidt-Küntzel A, Cornet A, Berzi P, Cambisano N, Ford C, Grisart B, Johnson D, Karim L, Simon P, Snell R, Spelman R, Wong J, Vilkki J, Georges M, Farnir F, Coppieters W: **Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition.** *Genetics* 2003, **163**:253–266.

29. Viitala S, Szyda J, Blott S, Schulman N, Lidauer M, Mäki-Tanila A, Georges M, Vilkki J: **The Role of the Bovine Growth Hormone Receptor and Prolactin Receptor Genes in Milk, Fat and Protein Production in Finnish Ayrshire Dairy Cattle.** *Genetics* 2006, **173**:2151–2164.

30. Cohen-Zinder M, Seroussi E, Larkin DM, Loor JJ, Everts-van der Wind A, Lee J-H, Drackley JK, Band MR, Hernandez AG, Shani M, Lewin HA, Weller JI, Ron M: **Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle.** *Genome Res* 2005, **15**:936–944.

31. Eberlein A, Takasuga A, Setoguchi K, Pfuhl R, Flisikowski K, Fries R, Klopp N, Fürbass R, Weikard R, Kühn C: **Dissection of Genetic Factors Modulating Fetal Growth in Cattle Indicates a Substantial Role of the Non-SMC Condensin I Complex, Subunit G (NCAPG) Gene.** *Genetics* 2009, **183**:951–964.

32. Karim L, Takeda H, Lin L, Druet T, Arias JAC, Baurain D, Cambisano N, Davis SR, Farnir F, Grisart B, Harris BL, Keehan MD, Littlejohn MD, Spelman RJ, Georges M, Coppieters W: **Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature.** *Nat Genet* 2011, **43**:405–413.

33. Littlejohn M, Grala T, Sanders K, Walker C, Waghorn G, Macdonald K, Coppieters W, Georges M, Spelman R, Hillerton E, Davis S, Snell R: **Genetic variation in PLAG1 associates with early life body weight and peripubertal weight and growth in Bos taurus.** *Anim Genet* 2012, **43**:591–594.

34. Pausch H, Aigner B, Emmerling R, Edel C, Götz K-U, Fries R: **Imputation of high-density genotypes in the Fleckvieh cattle population.** *Genet Sel Evol* 2013, **45**:3.

35. Pausch H, Wang X, Jung S, Krogmeier D, Edel C, Emmerling R, Götz K-U, Fries R: **Identification of QTL for UV-Protective Eye Area Pigmentation in Cattle by Progeny Phenotyping and Genome-Wide Association Analysis.** *PLoS One* 2012, **7**:e36346.

36. Eck SH, Benet-Pagès A, Flisikowski K, Meitinger T, Fries R, Strom TM: **Whole genome sequencing of a single Bos taurus animal for single nucleotide polymorphism discovery.** *Genome Biol* 2009, **10**:R82.

37. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**:1754–1760.

38. 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.

39. Picard: publicly available from http://picard.sourceforge.net.

40. Florea L, Souvorov A, Kalbfleisch TS, Salzberg SL: **Genome assembly has a major impact on gene content: A comparison of annotation in two Bos taurus assemblies.** *PLoS One* 2011, **6**:e21400.

41. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations.** *Nat Meth* 2010, **7**:248–249.

42. Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J: **The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species.** *Nucl Acids Res* 2001, **29**:159–164.

43. Gremme G, Brendel V, Sparks ME, Kurtz S: **Engineering a software tool for gene structure prediction in higher organisms.** *Inf Softw Technol* 2005, **47**:965–978.

44. Lewis SE, Searle SMJ, Harris N, Gibson M, Iyer V, Richter J, Wiel C, Bayraktaroglu L, Birney E, Crosby MA, Kaminker JS, Matthews BB, Prochnik SE, Smithy CD, Tupy JL, Rubin GM, Misra S, Mungall CJ, Clamp ME: **Apollo: a sequence annotation editor.** *Genome Biol* 2002, **3**:12.