Proposal of strategic investment for ICAR

Proponent: Interbull Subcommittee
Date: July 17, 2014
Subject: International Genotype Exchange Platform (GENOEX)
Parentage SNP exchange service (PSE)

BACKGROUND

Interbull service users have manifested interest in developing an international platform for genomic data exchange to enable Interbull to provide customized services to the current users and also to different players that are not directly involved with genetic evaluations but with genomic data. In order to identify clearly the existing demand, a survey among Interbull service users and collaborators was conducted in August 2012 and a total of 48 responses from 30 different countries were received. Considering the feedback received and the technical aspects, the Interbull Centre recommended adopting the BC|SNPmax data management system, by Biocomputing Platforms, as the database and SNP handling tool. This option was corroborated by an expert group invited by the Interbull Steering Committee. The Interbull Data Exchange Area (IDEA) will continue to be developed in parallel to manage pedigrees, national and international genetic evaluation data, and communication between the two databases will be developed. The objectives of the GENOEX proposal are: a. establishing the infrastructure necessary for international cooperation based on SNP data; b. optimizing customer investments in genotyping by avoiding duplication; c. establishing standard protocols for genomic data exchange; d. becoming the international source of bovine parentage SNPs; e. facilitating multilateral SNP data exchange by establishing a common repository and customer driven access rules; and f. providing affordable genomic data storage for small populations. The services to be provided through the implementation of GENOEX platform at the Interbull Centre are differentiated into three categories: parentage SNP exchange service (PSE), genomic data exchange service (GDE) and customized genomic repository service (CGR).

The GENOEX proposal was presented (Appendix I) and extensively discussed during the 2014 ICAR/Interbull Conference in Berlin, Germany, and considering the feedback received the Interbull Steering Committee has decided to go ahead with an immediate implementation of the PSE service and leave the GDE and CGR services for a second stage. Given that parentage verification and discovery represents an area of intersection between Interbull and other ICAR working groups, the Interbull Steering Committee decided to solicit that the ICAR board consider using resources from the ICAR strategic investment fund to cover part of the necessary investments.

THE GENOEX PROPOSAL

The detailed description of the proposed infrastructure and associated services is given at Dürr et al. (2014)[1], which is reproduced in Appendix I.

PROPOSED INVESTMENT

It is proposed that ICAR invests a total of € 60 000, in three annual installments of € 20 000.

BUDGET

The full budget for implementing the GENOEX-PSE is given in Table 1.

Software expenditures

- A service contract with the company Biocomputing Platforms Ltd Oy, from Finland, for the provision of the software (database and processing) has been signed by the Interbull Centre, and the figures given in Table 1 follow the agreement literally regarding software expenditures.
- The agreement signed with BC Platforms exempts the Interbull Centre from paying for a database license and for training, expenses that are regularly included in their proposals.
- The IBM DB2 Universal Database is a version free of charge that is sufficient for the purposes of the GENOEX-PSE.
- After the 1st year, only a fee to cover maintenance, update and support of the software is charged.

Hardware expenditures

- Investments in hardware are depreciated over 36 months.
- A reserve for upgrading the system is the only expenditure budgeted for the 4th year.

Other expenditures

- Programmer/database administrator salary: the first year will require more dedication to set up the system and also to collaborate with the customization work.
- Geneticist salary: besides resolving technical issues referring to the genotypes, this staff will provide customer services as well. It is expected that after the three implementation years the demand will be reduced.
- Overhead covers all other administrative and infrastructural costs associated to the services.

Income

- The Department of Animal Breeding and Genetics of SLU will invest € 20 000 in the first year.
- It is proposed that the service fees should generate approximately € 30 000 per year in the first three years to add up to the investments from ICAR and SLU and cover the total expenditures.
- From the 4th year onwards the service fees will have to cover all the expenses with the GENOEX-PSE.

[1] Dürr, J., Jorjani, H., Reents, R. 2014. International Genotype Exchange Platform (GENOEX). Proc. ICAR/Interbull Conference, Berlin, Germany, May 19-23, 2014. 10 p.

- Since Interbull has built financial reserves over the past two years, these should be used to cover eventual negative balance in the projects, as it is estimated to happen in the first year. Future surpluses from the GENOEX-PSE will be used to reset the reserves.

Table 1 – Budget for the implementation of the Parentage SNP exchange service (PSE) as part of the International Genotype Exchange Platform (GENOEX).

	Budget (€)			
Expenditures	**1st Year**	**2nd Year**	**3rd Year**	**4th Year**
Software:				
BC|OS without data amount limitations	14839	0	0	0
BC|DATA database license	0	0	0	0
Server installation and setup work	1845	0	0	0
Training	0	0	0	0
Customization work	8121	0	0	0
Maintenance, update and support fees	0	4592	4592	4592
IBM DB2 Universal Database	0	0	0	0
Sofware expenditures:	**24805**	**4592**	**4592**	**4592**
Hardware:				
Database server	4579	4579	4579	0
Processing server	1844	1844	1844	0
Upgrades	0	0	0	1659
Hardware expenditures:	**6423**	**6423**	**6423**	**1659**
Other:				
Programmer/DB Admin (Yr1: 33%; Yr2-3: 20%)	23500	14000	14000	14000
Scientist (20%)	15000	15000	15000	7500
Overhead (33% of salaries)	12705	9570	9570	7095
Other expenditures:	**51205**	**38570**	**38570**	**28595**
Total expenditures:	**82433**	**49585**	**49585**	**34846**
Income				
ICAR	20000	20000	20000	0
SLU	20000	0	0	0
Interbull Centre	13000	0	0	0
Service fees	29433	29585	29585	34846
Total income:	**82433**	**49585**	**49585**	**34846**
Balance:	**0**	**0**	**0**	**0**

International Genotype Exchange Platform (GENOEX)

J. W. Dürr[1], H. Jorjani[1] & R. Reents[2]

[1]Interbull Centre, Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, PO Box 7023, 750 07 Uppsala, Sweden
[2]Vereinigte Informationssysteme Tierhaltung w.V., Heideweg 1, D - 27283 Verden, Germany

Abstract

Interbull service users have manifested interest in developing an international platform for genomic data exchange to enable Interbull to provide customized services to the current users and also to different players that are not directly involved with genetic evaluations but with genomic data. In order to identify clearly the existing demand, a survey among Interbull service users and collaborators was conducted in August 2012 and a total of 48 responses from 30 different countries were received. Considering the feedback received and the technical aspects, the Interbull Centre recommended adopting the BC|SNPmax data management system, by Biocomputing Platforms, as the database and SNP handling tool. This option was corroborated by an expert group invited by the Interbull Steering Committee. The Interbull Data Exchange Area (IDEA) will continue to be developed in parallel to manage pedigrees, national and international genetic evaluation data, and communication between the two databases will be developed. The objectives of the GENOEX proposal are: a. establishing the infrastructure necessary for international cooperation based on SNP data; b. optimizing customer investments in genotyping by avoiding duplication; c. establishing standard protocols for genomic data exchange; d. becoming the international source of bovine parentage SNPs; e. facilitating multilateral SNP data exchange by establishing a common repository and customer driven access rules; and f. providing affordable genomic data storage for small populations. The services to be provided through the implementation of GENOEX platform at the Interbull Centre are differentiated into three categories: parentage SNP exchange service (PSE), genomic data exchange service (GDE) and customized genomic repository service (CGR). A step-wise implementation process will be adopted, starting by PSE followed by GDE and CGR.

Keywords: single nucleotide polymorphism, parentage verification, Interbull, database

Introduction

The use of genomic information for genetic evaluation of cattle populations has revolutionized the animal breeding business. As a consequence, Interbull service users and collaborators have manifested interest in developing an international platform for genomic data exchange within Interbull which would benefit dairy cattle breeding worldwide by (Dürr & Philipsson, 2012):

- reducing costs and optimizing investments on genotyping bovine animals;
- improving reference populations for prediction of genomically enhanced genetic merit, especially for low heritability health and functional traits, such as somatic cell count, mastitis, calving difficulty, longevity and female fertility;
- making it possible to screen large populations for recessive alleles detection;

- maintaining a worldwide parentage verification data base, using the SNP based methods that are about to be officially implemented by ISAG and ICAR; and
- using the genomic data to study diversity within the bovine populations in a more complete way than is possible with the methods based on pedigree information only.

In order to clearly identify the existing demands, a survey among Interbull service users and collaborators was conducted in August 2012 and a total of 48 responses from 30 different countries were received. The main conclusions from the survey were:

- There was clear interest from the customers (immediate establishment):
  - Common list of animals already genotyped in each population
  - Additional information on AI bulls (recessive alleles, for instance)
  - Exchange platform for parentage SNPs
- There was sufficient interest from customers (worth to invest):
  - Exchange platform for low density, middle density and high density SNP arrays
  - Inbreeding monitoring service
- There is clear interest from a good number of customers (customized demand):
  - National genomic data storage
  - Exchange platform for full sequences
- Depending on the establishment of a genomic data base (2nd phase):
  - Imputations
  - Multi-country evaluations
  - Genomic tools
- Different groups demanded different types of services and rejection by some should not prevent the others to be offered a service
- Countries with less advanced national genomic programs expect more from Interbull and are more prone to use the services investigated
- The pricing policy must reflect the multitude of services to be provided
- Legal implications of sharing genomic information through Interbull need to be clarified
- Any initiative in this field should include the participation of organizations that are not among the usual Interbull service users but that are directly affected by the implementation of such platform. Interbull has no interest to compete with the national organizations responsible for animal genotyping or parentage verification, but rather offer means to add value to their business.

This paper describes in detail the Interbull Centre proposal of implementing the International Genotype Exchange Platform (GENOEX) to attend the expectations of the Interbull community.

Objectives

The GENOEX project aims to add value to its future users by:

- establishing the infrastructure necessary for international cooperation based on SNP data;
- optimizing customer investments in genotyping by avoiding duplication;
- establishing standard protocols for genomic data exchange;
- becoming the international source of bovine parentage SNPs for national parentage verification/discovery service providers;

- facilitating multilateral SNP data exchange by forming a common repository and customer driven access rules; and
- providing affordable genomic data storage for small populations.

Choice of tools

Three options were initially considered to establish the GENOEX platform: in house development, commercial software and adopting an existing exchange platform e.g. the one developed by the Irish Cattle Breeding Federation (ICBF) known as IGenoP. The main favourable and unfavourable features of the three options are listed in Table 1. Considering the feedback received and the technical aspects involved, the Interbull Centre recommended adopting the BC|SNPmax data management system, by Biocomputing Platforms, as the database and SNP handling tool. This option was corroborated by an expert group invited by the Interbull Steering Committee. The Interbull Data Exchange Area (IDEA) will continue to be developed in parallel to manage pedigrees, national and international genetic evaluation data, and communication between the two databases will be developed.

Table 1. Comparison between three options of genomic data base implementations considered for GENOEX.

Option	Pros	Cons
In house development	• Lower initial cost • Built in integration with Interbull Data Echange Area (IDEA) • Customized solution • Possible to adopt open source solutions, such as PostgreSQL	• Complexity of the task • Development time • Lack of in house resources specialized in bioinformatics • Large number of other projects competing for resources
iGenoP	• Lower initial cost • Already operational • Developed based on similar needs	• Uses Oracle (not available at SLU, and license adds a significant cost) • Transposing the Oracle tables/codes into PostgresSQL would require a significant investment • ICBF is not a software company and further development would need to be in house • ICBF manifested no interest to host the services on behalf of Interbull
BC|SNPmax data management	• Strong technical background from presenting solutions to human genomics • Already in use for bovine genomics	• Higher initial cost • Uses IBM DB2, while IDEA uses PostgreSQL, but

system, by Biocomputing Platforms[1]	• Well established company with clients among Interbull customers • Offers solutions/tools highly sophisticated which would be difficult to develop in house • Tailored for multiple users with customized settings • Faster to implement • Benefits promoted by BC Platforms: o A ready-made system with fast implementation and deployment o All research data is securely stored, managed, shared, and analysed in one database o The data is always up-to-date, with logs and audit trails on data entries and modifications o Easy entering of data and sharing of data with chosen users and/or groups o A web based system enables remote access o More than 30 academic analysis tools supported, including various mendelian error check, linkage, imputing and GWAS tools o A cost effective way to outsource technical database maintenance, but keeping the data management in the hands of the research group o Tools to streamline whole data workflow and keep up with project status, leaving more time to actual research o Efficient, parallel data analysis workflow makes data analysis highly efficient	the license can be obtained at low cost • Liability of depending on a commercial software for service operations

[1] Biocomputing Platforms Ltd Oy, Tekniikantie 14, FI-02150 Espoo, Finland (www.bcplatforms.com).

Proposed service categories

The services to be provided through the implementation of GENOEX at the Interbull Centre are differentiated into three categories: parentage SNP exchange service, genomic data exchange service and customized genomic repository service. Figure 1 presents how these different services are structured, the connections between the different database partitions and access levels for different customers.

Parentage SNP exchange service (PSE)

Recent standards for parentage verification adopted by ISAG and corroborated by ICAR created a worldwide demand for a common repository of sets of SNPs to be used for both parentage verification and discovery. Since in most countries these activity is not performed by the usual Interbull customers (genetic evaluation bodies) but by organizations that currently are only indirectly related to Interbull (e.g. herd books, breed associations), the proposed service needs necessarily to include these potential new partners and also address their needs and priorities in order to become a valuable Interbull activity. One fundamental premise is that Interbull will not become a parentage verification provider and competitor of the already established business, but rather a facilitator of data exchange among them.

ICAR member representatives will be requested to inform Interbull which organizations in their respective countries are responsible for the official parentage verification services. These organizations will then be pre-registered on an ICAR Parentage Verification Directory (to be created) and offered to join the Parentage SNP Exchange Service. Among these, two categories of users will be defined:

- Interbull user: organization that is already participating in Interbull international genetic evaluation services and therefore has already access to the Interbull Data Exchange Area (IDEA) which hosts the international pedigree and phenotypic information used for international genetic evaluations.
- External user: organization that is an associate or full member of ICAR or represents a full member of ICAR and that is responsible for official parentage verification services in its own coverage area. No access to IDEA will be granted to external users.

In either case, only organizations that possess suitable data (SNPs, microsatellites) and agree to contribute to the common pool of data will be entitled to enroll in the services.

Modus operandi

This module will handle three data types:

- Parentage confirmation SNPs (~100)
- Parentage discovery SNPs (~400)
- Parentage confirmation microsatellite markers

The principle of the service is a cooperative effort to facilitate parentage verification worldwide by providing access to SNP and microsatellite information on foreign animals not available at the national databases. Provided that the user is by default officially responsible for parentage verification services within its respective country-breed, it will be expected that each participant contributes with data to the common repository of genotypes in a continuous basis (especially sires and dams). Likewise, each user will be granted access to the full directory of parentage SNP and microsatellite data available. As a cooperative process, the service charges will not be based on the usage rate, but rather be defined as a common annual subscription fee.

In order to ensure transparency, an updated log of the user activities will be made available to all users for monitoring uploads and downloads. Users of the Parentage SNP Exchange Service will use a secure web interface for both uploading and downloading of parentage data.

Consistency checks will be performed by an uploading filter, as indicated in Figure 1, and illustrative examples of rules are given in Table 2.

Table 2. Rules for uploading and downloading data for the Parentage SNP Exchange Service.

Type of rule	**Example**
Data accessibility	Only official national parentage verification service suppliers will be allowed to join the service
Uploading consistency checks	Correct file formats? Animal already in the Interbull pedigree (IDEA)? Animal parentage has been confirmed/verified previously? Animal already has parentage SNPs in the Genoex DB?

Alternative approach

There is an alternative approach being proposed by some collaborators that want to participate in an international cooperation for parentage verification but have internal impediments to share the actual parentage SNP data. Describing such proposition is out of the scope of this document, but it generally consists on developing software that would communicate with the different national databases and perform the parentage verification on behalf of the users without exchange of SNP data among them. The focus of the service would therefore be on the results of parentage verification only. Such development would not be carried out by the Interbull Centre and possibly the services would also be outsourced to the developer, given the complexity involved in the maintenance of the system. The expert group consulted by the Interbull Steering Committee did not recommend the adoption of this alternative by Interbull due to the intricacy involved in this kind of development that needs to interact with a multitude of different national data repositories and firewall policies.

Genomic data exchange service (GDE)

Genomic evaluations require assembling significantly large reference populations from which both SNP and phenotypic data has to be available. This requirement has stimulated breeding organizations to exchange marker genotypes and share reference populations avoiding multiple genotyping of the same individuals. Bilateral exchanges may become inefficient and cumbersome if clear standards for data sharing are not established and multiple partners are involved. For this reason, Interbull customers have demanded that an SNP data exchange platform is established to facilitate the exchanges.

Only organizations that are already participating in Interbull international genetic evaluation services (Interbull users) and therefore have access to the Interbull Data Exchange Area (IDEA) which hosts the international pedigree and phenotypic information used for international genetic evaluations will be offered this specific service. The Interbull users that enroll in the GDE service will fully benefit from the PSE service as well if they are listed in the ICAR Parentage Verification Directory.

Modus operandi

Data types:

- Parentage confirmation SNPs (~100)
- Parentage discovery SNPs (~400)
- Parentage confirmation microsatellite markers
- Low density SNP arrays
- Middle density SNP arrays
- High density SNP arrays

The GDE service will be based on three fundamental principles:

1. Ownership of genotypes belongs to the organization supplying the genotype in the first place and consequently the control of the access rules for users that may have access to the information.
2. Users are charged proportionally to the respective downloading activity.
3. Users are granted credit proportionally to the respective uploading activity.

Each user of the GDE service will have full access to its own marker genotypes plus those genotypes obtained by exchange within the system. The access permissions will be defined by the genotype owners directly in the user interface, without interference of the Interbull Centre. Likewise, genotype requests will be managed by the user console through a message system. Since the Interbull Centre will not be directly involved in the exchanges but in maintaining the database tools, a general service contract will be signed with each user. The legal framework of specific bilateral or multilateral genotype exchanges will have to be established between the parties involved. Users of the Genomic Data Exchange Service will use a secure web interface for both uploading and downloading of data. Consistency checks and data accessibility rules will be performed by filters, as indicated in Figure 1, and examples of rules are given in Table 3.

Table 3. Rules for uploading and downloading data for the Genomic Data Exchange Service.

Type of rule	Example
Data accessibility	Only Interbull users can subscribe to the services
	Permission from owner granted?
Uploading consistency checks	Correct file formats?
	Animal already in the Interbull pedigree (IDEA)?
	Animal already has marker genotype(s) in the Genoex DB?

Customized genomic repository service (CGR)

Genomic data is significantly more complex and space demanding than the conventional performance and pedigree records that Interbull customers have been dealing with in the past decades. This complexity is challenging the national databases not only in terms of hardware dimensions (hard disk, memory and processing capacity) but also in terms of the necessary tools to keep up with the multitude of SNP arrays available in the market. Storing genomic data in a ready-to-use way is not a trivial operation especially when human and financial resources are limited. For this reason, the infrastructure to be built at the Interbull Centre is an opportunity for Interbull customers to store and process their genomic information using exclusive partitions of the GENOEX database through remote access. In other words, national evaluation units would outsource the database services to the Interbull Centre, but keep exclusive access to the data.

Given that these customized repositories are built completely independent from each other, this opportunity can eventually be extended to different types of users within the ICAR framework.

Customer characterization:

- Interbull user: organizations that are already participating in Interbull international genetic evaluation services and therefore have access to the Interbull Data Exchange Area (IDEA) which hosts the international pedigree and phenotypic information used for international genetic evaluations. Enrollment in the Customized Genomic Repository Service does not include access to the other service categories described in this proposal.
- External user: organization that is an associate or full member of ICAR or represents a full member of ICAR. No access to IDEA will be granted to external users.

Modus operandi

Data types:

- Parentage confirmation SNPs (~100)
- Parentage discovery SNPs (~400)
- Parentage confirmation microsatellite markers
- Low density SNP arrays
- Middle density SNP arrays
- High density SNP arrays
- Sequencing data

This service involves storing the national genomic data in an exclusive area of the Genoex DB servers and providing also processing capacity of Interbull crunchers for the users.
The CGR service will also use the BC|SNPmax data management system, which offers the following extra features (besides the benefits included on Table 1):

- Extremely fast data writing to SQL database, or to compressed binary datasets
- Tools for querying research data
- Automated data conversion tools
- Easy and intuitive web-browser interface for different genetic analysis programs
- Interfaces for R, SAS, and other scripting languages
- In-built queuing system and automated data analysis job segmentation for running heavy genetic analysis to be run in parallel
- Optional support for cloud computing technologies
- Automated log-file functionality for data auditing
- Database on-line back-up (back-up while database is in use)

Users of the CGR service will use a secure web interface for accessing BC|SNPmax and will be fully responsible for data maintenance.

Implementation

The initial investments need to cover the BC|SNPmax installation and dedicated servers to host and operate the database. Integrating the BC|SNPmax and the IDEA platforms as well as developing data filters and the user interfaces will be done in collaboration with BC Platforms. Infrastructure grants will cover the initial investments and service fees will be charged to cover direct costs associated to the services provided.

A stepwise implementation process will be carried out, starting by the PSE service and following with the GDE service. The CGR modality will be developed according to the demand. Table 4 presents the proposed implementation calendar.

Table 4. Proposed implementation calendar for GENEX services.

Period	Activity
Summer 2014	BC|SNPmax license purchase + installation
Summer – Fall 2014	Development of the parentage SNP exchange (PSE) service interface
Winter 2014/15	Launching of the PSE service
Spring – Summer 2015	Development of the genotype data exchange (GDE) service interface and integration with IDEA
Winter 2015/16	Launching of the GDE service
2016	Development of the customized genomic repository (CGR) service

Final remarks

The GENOEX proposal represents a unique opportunity for ICAR/Interbull to provide a much needed international platform for genomic information exchange. Demand has been identified and the means presented here show that it is not only feasible, but also affordable. The projected service fees are remarkably lower than isolated investments necessary for genotyping and genomic data storage/handling. As a cooperative effort that would involve only those customers interested in benefiting from the proposed platform, it does not prevent other Interbull customers to continue developing their own strategies and infrastructure. As it is the case for all Interbull activities, there is no intention to replace or compete with national organizations; on the contrary, GENOEX is conceived to offer auxiliary tools for the national expertise to develop their own programs more efficiently.

List of references

Dürr, J.W. & J. Philipsson. 2012. International cooperation: The pathway for cattle genomics. Animal Frontiers, 2(1): 16-21.

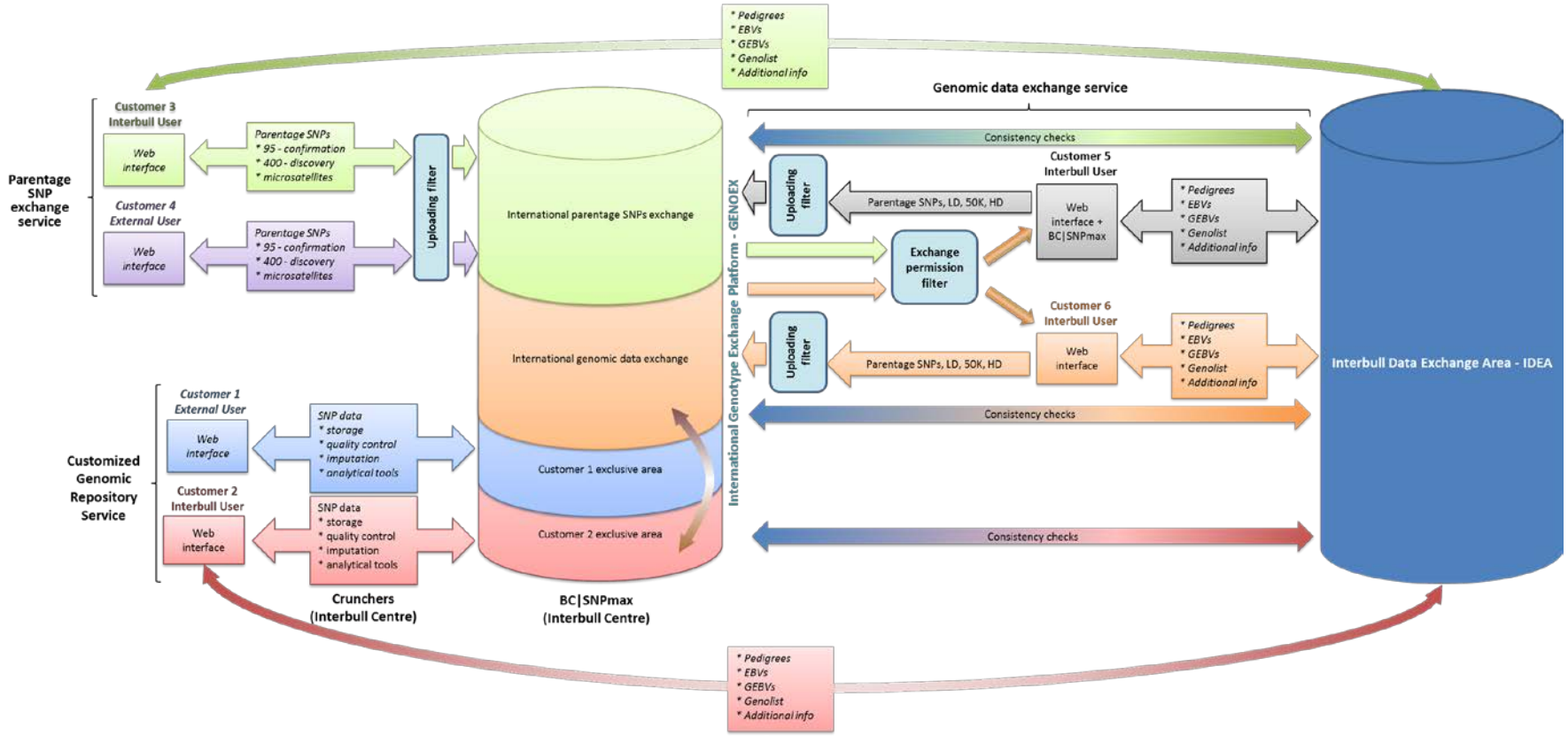

Figure 1. Graphical representation of the services to be provided through the implementation of the international genotype exchange platform (Genoex) at the Interbull Centre.