Journal of Genetics and Molecular Biology
Vol. 17, No. 4, 204-210, December 1, 2006

Qualiseq: Quality Genomic Sequence Retrieval

Yu-Chang Cheng, Tz-Chao Lin, Kuo-His Lee, Yan-Hau Chen, Ming-Fang Tsai, Yi-Jung Lin, and Adam Yao

National Genotyping Center, Institute of Biomedical Sciences, Academia Sinica, Taipei, 115, Taiwan

Genomic sequence analysis starts from a sequence with poor quality such as mis-orientation, incorrect location or contig order can hardly produce fruitful results in final assays. Unfortunately, more than often the circumstance is not found early enough until a lot of time and effort has been invested. To alleviate this situation, we have developed QualiSeq, a web application with a friendly interface providing information on sequence quality of a genomic region so that researchers can easily download much more reliable sequences for further analysis and assay design.

Key Words*: Sequence, download, web, alignment*

Introduction

Downloading a sequence from online public databases provided by NCBI (National Center for Biotechnology Information) [1], Ensembl [2] and UCSC for sequence analyses is a very common step in genomic research. To assure the sequence quality, we need to perform a series of examinations before knowing if the sequence is identical among major sources as well as in correct orientation and assembly, containing no long repeats, etc. Usually, a researcher would download sequences only from a single database that he/she is familiar with. With database version keeps changing and is not always synchronized among various sources, poor quality sequences can be retrieved unexpectedly. To avoid the problem created by version difference, a researcher needs to retrieve sequences from various sources and compare them. Unfortunately, retrieving a sequence is not an intuitive and obvious procedure for any of the above three web sites not to mention that they all have different procedures. Besides the above consistency check, the orientation, assembly, and repeats all need to be looked over closely. Consequently, it takes a lot of operations to finish the examinations. What's more, these operations are also very error-prone by manual.

As a result, we have created QualiSeq

*Corresponding author: Dr Adam Yao, Inst of Biomed Sciences, Academia Sinica, 128 Academia Rd, Sec 2, Nankang, Taipei 115, Taiwan
Availability: http://genepipe.ngc.sinica.edu.tw/qualiseq/
Contact: adam@ibms.sinica.edu.tw

This article is available on line at: http://abrc.sinica.edu.tw/jgmb/

to help researchers to download quality sequences in a much easier way. The usage flow of QualiSeq is straightforward. First, researchers designate a genomic region, then click the “submit“ button. The update of progress and final results of the submitted query are shown right below the button in the same page including three tabbed subpages to display the current status of SNP information extraction, consistency check, and assambly check, respectively. If there is an error in any step, QauliSeq will stop doing the remaining steps (Figure 1).

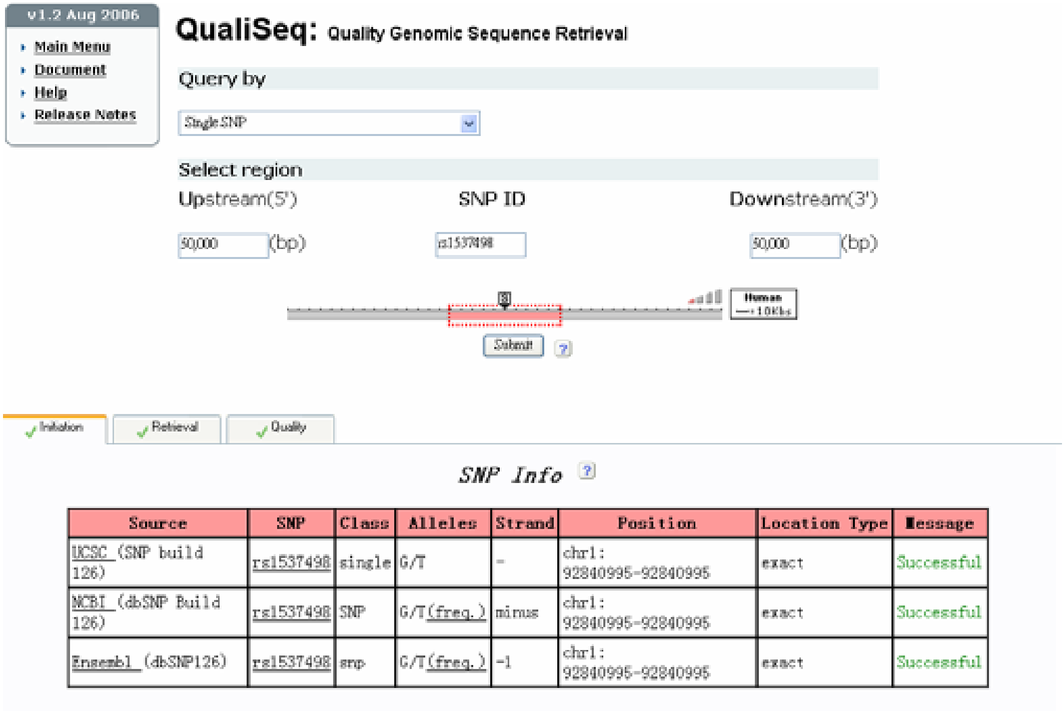
QualiSeq automates all procedures that are needed to assure the quality of retrieved sequences in order to reduce tedious and complicated operations into simple typing and clicks.

Implementation

Input:

QualiSeq provides various ways for researchers to define their interested regions. For example, a single marker (SNP or STRP) or a pair of markers plus its downstream and upstream can be used to define a retrieval region. Alternatively, physical positions can also be used to retrieve a sequence directly (Figure 2).

QualiSeq accepts many SNPs or STRPs as a batch query as well. Researchers type all interested makers ID or marker names in which they are interested to a text file and upload the file to QualiSeq. The maximum number of markers that QualiSeq allows to handle is 100 per batch query. And the maximum allowable length of a query range is 4 Mbps.



Source	SNP	Class	Alleles	Strand	Position	Location Type	Message
UCSC (SNP build 126)	rs1537498	single	G/T	-	chr1: 92840995-92840995	exact	Successful
NCBI (dbSNP Build 126)	rs1537498	SNP	G/T(freq.)	minus	chr1: 92840995-92840995	exact	Successful
Ensembl (dbSNP126)	rs1537498	snp	G/T(freq.)	-1	chr1: 92840995-92840995	exact	Successful

Fig. 1. The result page containing three tabbed subpages to display the current status of SNP information extraction, consistency check, and assambly check, respectively.

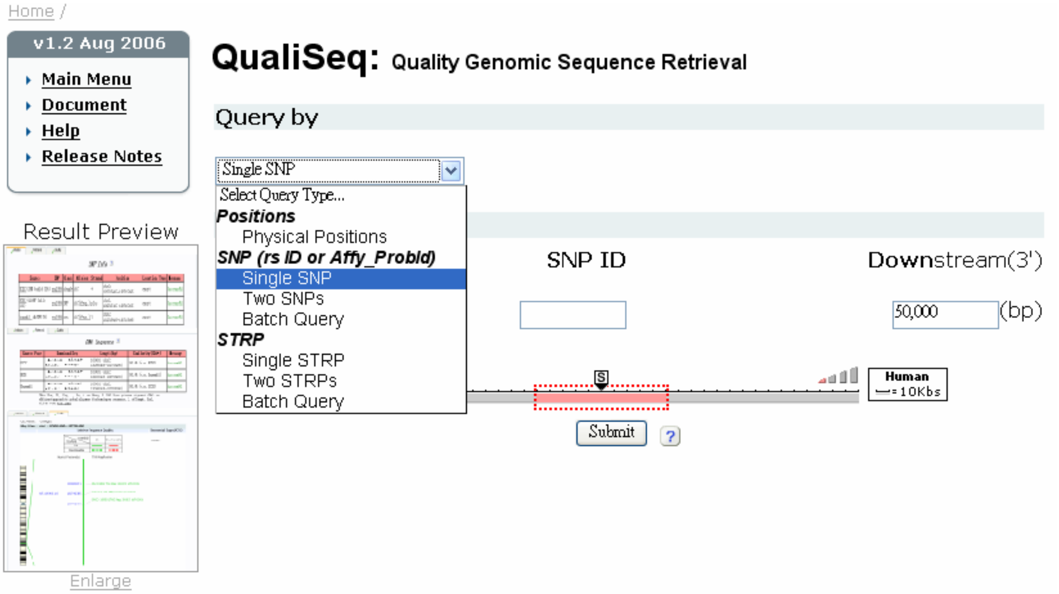
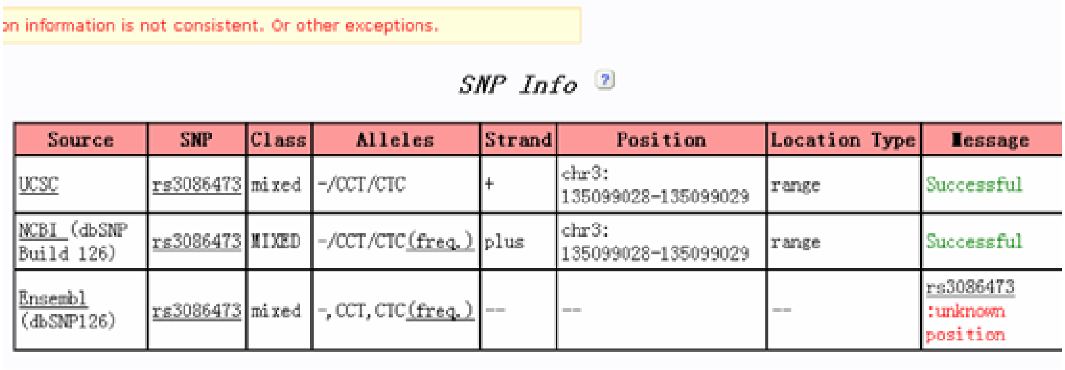


Fig. 2. The input page of QualiSeq.



on information is not consistent. Or other exceptions.

SNP Info ?

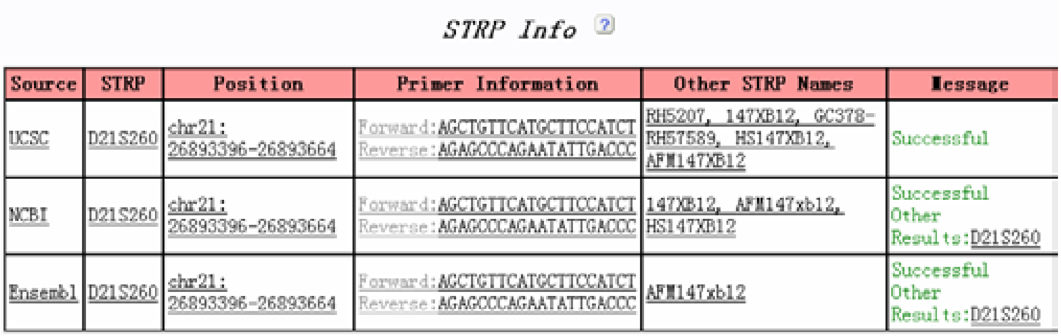
Source	SNP	Class	Alleles	Strand	Position	Location Type	Message
UCSC	rs3086473	mixed	-/CCT/CTC	+	chr3: 135099028-135099029	range	Successful
NCBI (dbSNP Build 126)	rs3086473	MIXED	-/CCT/CTC(freq.)	plus	chr3: 135099028-135099029	range	Successful
Ensembl (dbSNP126)	rs3086473	mixed	-, CCT, CTC(freq.)	--	--	--	rs3086473 :unknown position

Fig. 3. Inconsistent SNP information table presented on the Initiation subpage.

Initiation stage:

QualiSeq shows marker information extracted from NCBI, Ensembl and UCSC in tabulated form for easy comparison. Then QualiSeq matches the positions of input makers in genome. However, QualiSeq allows up to 5% difference of marker physical position value otherwise it terminates running the application. If markers are near the beginning of a chormosome, the accepted physical distance difference between each marker position would be smaller than the markers near the end of a chormosome. The reason is that maker position around the end of chormosome would have greater chance to be affected than makers around the beginning of chormosome by genome sequences update.

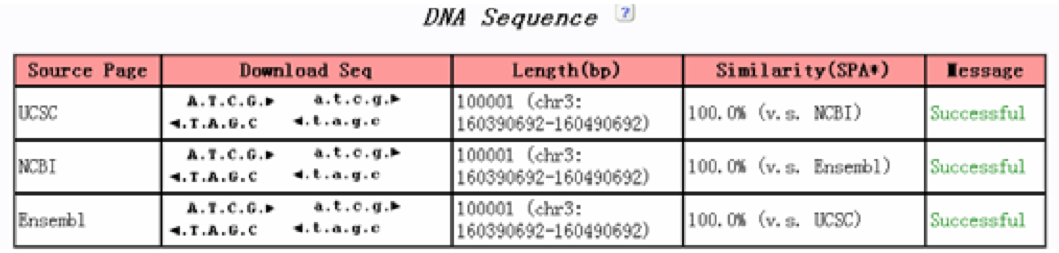
QualiSeq also presents information such as position, type, allele, primers, ...etc. in a table (Figure 3 and Figure 4). The following situations cause QualiSeq to halt on this stage: a maker with no sequence map, a maker with more than one position or



STRP Info

Source	STRP	Position	Primer Information	Other STRP Names	Message
UCSC	D21S260	chr21: 26893396-26893664	Forward:AGCTGTTCATGCTTCCATCT Reverse:AGAGCCCAGAATATTGACCC	RH5207, 147XB12, GC378-RH57589, HS147XB12, AFM147XB12	Successful
NCBI	D21S260	chr21: 26893396-26893664	Forward:AGCTGTTCATGCTTCCATCT Reverse:AGAGCCCAGAATATTGACCC	147XB12, AFM147xb12, HS147XB12	Successful Other Results:D21S260
Ensembl	D21S260	chr21: 26893396-26893664	Forward:AGCTGTTCATGCTTCCATCT Reverse:AGAGCCCAGAATATTGACCC	AFM147xb12	Successful Other Results:D21S260

Fig. 4. STRP information table presented on the Initiation subpage.



DNA Sequence

Source Page	Download Seq	Length(bp)	Similarity(SPA*)	Message
UCSC	A.T.C.G.▸ a.t.c.g.▸ ◂.T.A.G.C ◂.t.a.g.c	100001 (chr3: 160390692-160490692)	100.0% (v.s. NCBI)	Successful
NCBI	A.T.C.G.▸ a.t.c.g.▸ ◂.T.A.G.C ◂.t.a.g.c	100001 (chr3: 160390692-160490692)	100.0% (v.s. Ensembl)	Successful
Ensembl	A.T.C.G.▸ a.t.c.g.▸ ◂.T.A.G.C ◂.t.a.g.c	100001 (chr3: 160390692-160490692)	100.0% (v.s. UCSC)	Successful

Figure5 QualiSeq results showing at the Retrieval subpage

no position listed from original website, and data retrieval problems caused by network problem.

Retrieval stage:

At the retrieval stage, QualiSeq samples a sequence from the same regions of three major genomic sequence sources, NCBI, UCSC and Ensembl to make sure the sequence in the region is consistent in the public domain.

To evaluate sequence consistency in the public domain, the retrieved three sequences are compared to one another for similarity check. Here QualiSeq uses SPA (Super Pairwise Alignment) tool [4] to score their similarity. Only when all sequences are identical or near identical (the threshold is 99.99% identity) will trigger the quality examination stage.

The sequences from three sources can be downloaded in both directions. In addition, researchers can choose if they want the downloaded sequence in capital letters or not (Figure 5).

Quality check stage

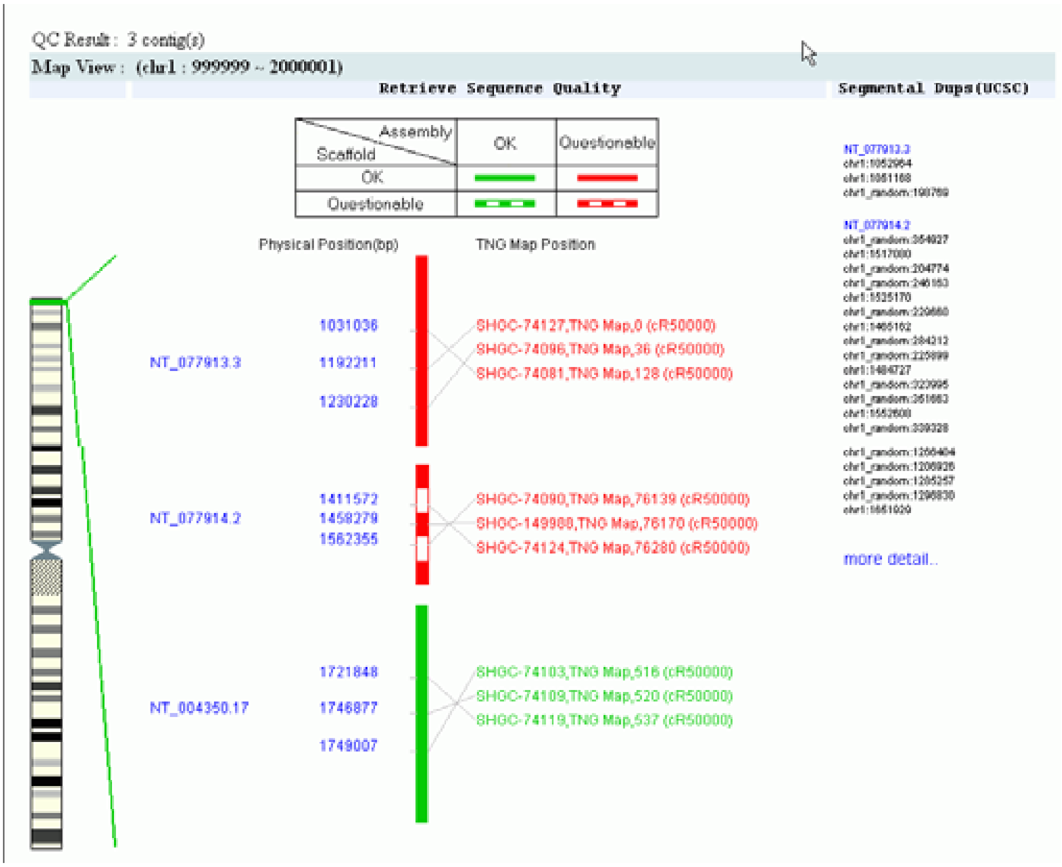
As we know that human genomic sequence map is built from many smaller contigs with an algorithm to arrange them in order, errors could be produced in this process. Therefore, we implement a program that uses other map to confirm genomic sequence map and integrate it into our sequence check mechanism.

Entering the quality examination stage, QualiSeq looks into the assembly quality if the order of the TNG markers on RH map is the same as on the contig map. The result is displayed in green color when assembly check is passed otherwise it is in red. When two or more contigs exist in a retrieved region, three TNG markers from each contig will be selected to examine if the contig order is correct or questionable. When contig order check is passed, a solid bar will be shown, and if not, a broken bar will be seen.

Because of the resolution of TNG makers on RH map, the distance in genomic sequence map between two disordered TNG markers should be greater than 60kbps to consider as real disordered TNG markers. Segmental duplications that come from UCSC database will be shown beside the contig graph. All these data are visualized as graphs and tables to help researchers have an idea of the sequence quality (Figure 6 and Figure 7). The update of human genome can be clearly shown by QualiSeq as Figure 6 and Figure 7. The NCBI genome build 36.2 has fewer contigs than build 35.1. It means that the coverage of human genome sequence is getting more complete. However, the quality check of retrieved genome sequence is still needed in the present day.

For the batch marker inputs, there is a Genome View figure on the top of the result page presenting to researchers the distribution of input markers in whole chromosomes and gene names near these markers. Researchers can click an item on the figure and the detail information described in previous sections would show up in below. Researchers also are able to download all sequences of their batch queries by one click on the top right icons in .CSV file format (Figure 8).

Conclusion



QC Result : 3 contig(s)

Map View : (chr1 : 999999 ~ 2000001)

Retrieve Sequence Quality

Scaffold \ Assembly	OK	Questionable
OK	🟩	🟥
Questionable	🟩	🟥

Segmental Dups(UCSC)

NT_077913.3
chr1:1052964
chr1:1061168
chr1_random:190769

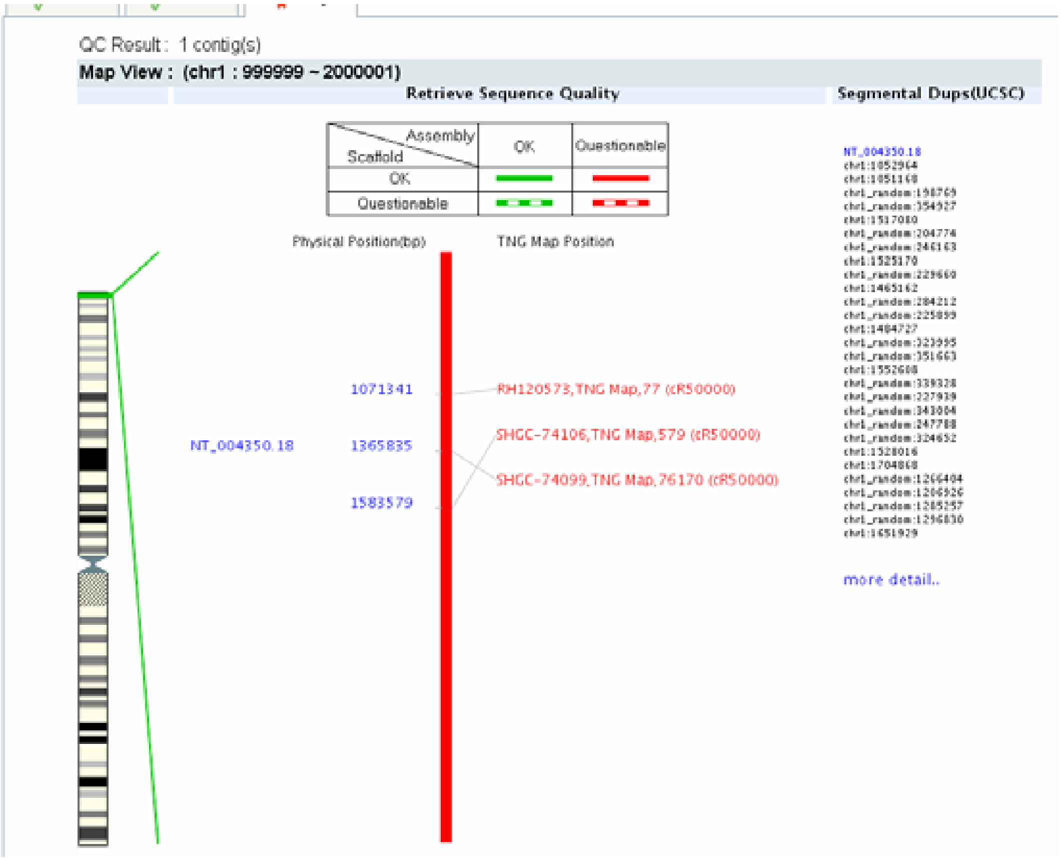
NT_077914.2
chr1_random:354027
chr1:1517000
chr1_random:204774
chr1_random:246183
chr1:1525170
chr1_random:229660
chr1:1465102
chr1_random:284212
chr1_random:229899
chr1:1484727
chr1_random:323995
chr1_random:251663
chr1:1552600
chr1_random:339328

chr1_random:1266404
chr1_random:1206926
chr1_random:1265257
chr1_random:1298830
chr1:1661029

more detail..

Physical Position(bp)
TNG Map Position
NT_077913.3
1031036
1192211
1230228
SHGC-74127,TNG Map,0 (cR50000)
SHGC-74096,TNG Map,36 (cR50000)
SHGC-74081,TNG Map,128 (cR50000)
NT_077914.2
1411572
1458279
1562355
SHGC-74090,TNG Map,76139 (cR50000)
SHGC-149988,TNG Map,76170 (cR50000)
SHGC-74124,TNG Map,76280 (cR50000)
NT_004350.17
1721848
1746877
1749007
SHGC-74103,TNG Map,516 (cR50000)
SHGC-74109,TNG Map,520 (cR50000)
SHGC-74119,TNG Map,537 (cR50000)

Fig. 6. Quality check result of a retrieved region with three contigs from NCBI human genome build 35.1 showing questionable assembly of contig NT_077913.3, questionable contig order and assembly of contig NT_077914.2, and a good quality contig NT004350.17.


1071341
NT_004350.18
1365835
1583579
RH120573,TNG Map,77 (cR50000)
SHGC-74106,TNG Map,579 (cR50000)
SHGC-74099,TNG Map,76170 (cR50000)
Physical Position(bp)
TNG Map Position

QC Result : 1 contig(s)

Map View : (chr1 : 999999 ~ 2000001)

Retrieve Sequence Quality

Scaffold \ Assembly	OK	Questionable
OK	🟩	🟥
Questionable	🟩	🟥

Segmental Dups(UCSC)

NT_004350.18
chr1:1052964
chr1:1051168
chr1_random:198769
chr1_random:354927
chr1:1517080
chr1_random:204774
chr1_random:246163
chr1:1525170
chr1_random:229660
chr1:1465162
chr1_random:284212
chr1_random:225899
chr1:1484727
chr1_random:323995
chr1_random:351663
chr1:1552608
chr1_random:339328
chr1_random:227939
chr1_random:343004
chr1_random:247788
chr1_random:324632
chr1:1328016
chr1:1704868
chr1_random:1266404
chr1_random:1206926
chr1_random:1285257
chr1_random:1296830
chr1:1651929

more detail..

Fig. 7. Quality result of the same region in build 36.2 showing one contig only but still has questionable assembly.

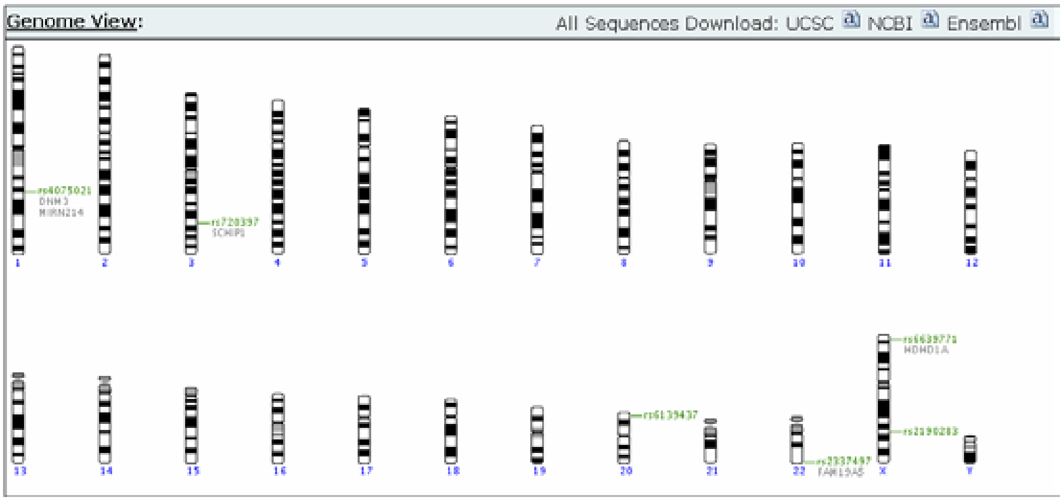


Fig. 8. The distribution view of input makers.

By utilizing automated processes to replace tedious and error-prone manual operations, QualiSeq makes quality sequences easy to obtain. The application is flexible in input selection and simple to use. Furthermore, it integrates three main genomic information sources and makes researchers to be aware of the differences and variances of genomic data among these data sources on the Internet. This feature is helpful for researchers who often solely use one data source and may miss the latest version of data published by other websites. With our visual presentation, the outcome is easily understood. QualiSeq should greatly help researchers focus on sequences with good reliability and make their downstream research more efficient and effective.

Acknowledgements

We appreciate support from Dr. Yuan-Tsong Chen and Dr. Jer-Yuarn Wu to realize the project. The project is funded by Academia Sinica Life Science Grant No. 40-19(GSK).

References

[1] Jenuth JP. The NCBI. Publicly available tools and resources on the Web. Methods Mol Biol, 2000; 132: 301-12.

[2] Birney E, Andrews D, Caccamo M, Chen Y, Clarke L., Coates, G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Flicek P, Graf S, Hammond M, Herrero J, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Kokocinski F, Kulesha E, London D, Longden I, Melsopp C, Meidl P, Overduin B, Parker A, Proctor G, Prlic A, Rae M, Rios D, Redmond S, Schuster M, Sealy I, Searle S, Severin J, Slater G, Smedley D, Smith J, Stabenau A, Stalker J, Trevanion S, Ureta-Vidal A, Vogel J, White S, Woodwark C, Hubbard TJ. :Ensembl 2006 Nucleic Acids Res. 2006 Jan 1; 34 Database issue: D556-D561.

[3] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler D: The human genome browser at UCSC. Genome Res, 2002; 12(6), 996-1006.

[4] Shen SY, Yang J, Yao A, and Hwang P: 2002 Super pairwise alignment (SPA): an efficient approach to global alignment for homologous sequences. J Comput Biol. 2002; 9(3):477-486s