Project Reconnect: A web-based application based on 384 SNPs to reconnect missing children with their parents in China

Luren Wang[1], Calvin Tjandra[1] and Jianyi Ren[2]

[1]Department of Computer Science, 1214 Amsterdam Ave, New York, NY 10027, [2]Department of Chemical Engineering, 500 W 120th St, New York, NY 10027.

Abstract

Motivation: Each year thousands of children in China were illegally taken from their parents or ended up as missing. In this study we selected a panel of 384 SNPs with high discrimination power over 14 ethnic groups in China which covers 96% Chinese population. We also established a database and web platform based on the selected SNPs that serves to reunion parents and children who are disconnected. The aim of this study is to keep pace with technology advancement and provide a SNP-based database, complementary to the classic STR system, to grant individuals a convenient access to their lost family members, interterritorially as well as internationally.

Availability: Project Reconnect is a free and publically available web tool which can be accessed at http://zhaohaizi.org.cn (waiting for Government administration approval)

Contact: lw2666@columbia.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Kidnapping and child trafficking is a serious social issue in China. Each year it is estimated 20,000~70,000 Chinese children were kidnapped (Bureau of Democracy, 2009; Neumann, 2007), with 13,000 rescued by police department of China(Liu, 2015). The trafficked children are sold to adoption families, domestic or international, turned into child labor, prostitution, or forced to engage in begging, theft and robbery. In some of the worst cases, children are permanently injured, amputated and turned into professional panhandler (Jiang and Sanchez-Barricarte, 2013). The effect of losing a child a devastating on a family.

Single Nucleoside Polymorphism(SNP) is a common generic variation among people. SNP panels have been developed as highly effective identification tools that pertain comparable precision to the classic short tandem repeats(STR) approach among a variety of populations (Augustinus, et al., 2015; Boonyarit, et al., 2014; Hwa, et al., 2016; Kim, et al., 2010; Pakstis, et al., 2010; Wei, et al., 2014; Zeng, et al., 2012). Compared to the STR method for paternity establishment, SNP system has the advantages of a significant lower mutation rate(Amorim and Pereira, 2005; Huang, et al., 2002; Reich, et al., 2002) . The rapid advancement (Kim and Misra, 2007) and decreasing costs of SNP genotyping techniques has lead to the availability of various commercial companies that provide individuals genotyping service at a highly competitive cost to STR system.

In this study we established a standard SNP panel for paternity testing in Chinese population and constructed an online database capable of storing and matching SNP genotypes. Given a small amount of initial capital investment, Project Reconnect could be brought live in a short amount of time with success

2 Methods

The project was developed in three different parts simultaneously. The first part was screen a panel of SNPs of high distinguishing resolution within Chinese population. The second part was the development of an efficient and thorough matching algorithm. The third part was the development of digital infrastructure in the form of a website and database for the project to be housed in.

2.1 SNP marker selection

SNP markers for paternity test in Chinese population were selected according to the following criteria: (1) SNPs were required to have minor allele frequency(MAF) between 0.35~0.45 in all of the 14 Chinese subpopulations, including Han, Dai, Daur, Hezhen, Lahu, Miaozu, Mongola, Naxi, Oroqen, She, Tu, Tujia, Xibo, Yizu . The MAF threshold was established based on two major considerations: high heterozygosities is necessary to obtain a high distinguishing power; however, SNPs with MAF approaching 0.5 are more prone to experiment errors (Itsik Pe'er, personal communication, April 25, 2016); (2) The SNPs are in Hardy-Weinberg equilibrium (HWE), to ensure within-locus independence of alleles; (3) The SNPs are in linkage equilibrium, to ensure between locus independence of the alleles; (4) SNPs are not associated to diseases, to avoid including sensitive information of users; (5) The SNPs selected were located outside the common copy number variance(CNV) regions of each

chromosome, to enhance performance of SNP genotyping experiment; (6) SNPs were reported by multiple mainstream genotyping service providers, to facilitate future users an easy data acquisition from various third party genotyping service providers.

SNPs were selected as the overlap of SNPs reported by 23andme, Ancestry, deCODEme, FamilyTree and Wegene. To screen candidate SNPs we used allele frequency data from Hapmap Phase III for CHB population (Gibbs, et al., 2003), 1000 Genomes Phase III for CHS (Consortium, 2015), HGDP-CEPH Human Genome Diversity Cell Line Panel for Dai, Daur, Hezhen ,Lahu, Miaozu, Mongola, Naxi, Oroqen, She, Tu, Tujia, Xibo, Yizu. SNPs frequencies were evaluated either by a R script or by SPSmart (Amigo, et al., 2008) and all chosen SNPs have a MAF between 0.35~0.45 for all 14 subpopulations. The SNPs associated to diseases according to DisGeNET v3.0 (Piñero, et al., 2015) were excluded. Chi-square tests were performed to determine Hardy-Weinberg equilibrium and alleles with $\chi^2 > 3.841$ were eliminated. SNPs locate in areas of CNV frequency > 1% according to DECIPHER Database (Firth, et al., 2009) were excluded. r^2 for Linkage equilibrium were calculated by LDlink (Machiela and Chanock, 2015). The mean, SD and maximum value of all within-chromosome pairwise r^2 for 384 SNPs are 0.0023, 0.0030, 0.017. A full list of the 384 SNPs can be found in supplementary table 1. Source code for the SNP screening can be found at https://github.com/Jianyi-Ren/ProjectReconnect.

2.2 Algorithm Development

The algorithm design for Project Reconnect was focused upon ways to quickly and efficiently compare large amounts of SNP sequences and finding matches among these sequences indicating DNA inheritance by descent. Initial development was focused upon modifying GERMLINE, an algorithm designed by Professor Pe'er at Columbia University. A modified GERMLINE with less features was under development in Python. However, research into the amount of SNP positions required to identify parent child relationships showed that the dynamic programming approach was not necessary for such short SNP sequences and this algorithm was abandoned in favor of a simpler matching algorithm. The final requirements for such an algorithm was rapid processing of a large number of short SNP sequences in order to find the number of matching SNP positions between one sample and the rest of the database.

2.3 Digital Infrastructure

We incorporated our core algorithm into a web framework for a working site. Our digital infrastructure is modeled after the traditional full stack model.

For our backend database, we used Postgresql. Postgresql is regarded as a mature and highly capable database and the most SQL compliant. For these reasons, we chose Postgresql over MySQL as Postgresql also contains JSON datatype which helps speedier prototyping without data migration. However, during development, we used SQLite as it was easier to work with, lighter, and most importantly, we can use version control.

Flask served as our web framework. There are several other frameworks we have considered including Django and Bottle. However, Flask was considered to be light enough for rapid development and still contain powerful features such as a good html templating language (Jinja2) and blueprints.

On the UI end, we used Bootstrap. It was easy to create very nice user interfaces using Bootstrap templates. Bootstrap contains ready-made UI elements such as buttons as well as an extremely useful grid-system. These allowed us to rapidly create a professional looking site.

Our web application followed the traditional MVC design pattern, where the user interacts with the views, which updates the models, and models subsequently updating the views. Our file structure reflects this design pattern. The views folder contains the url routing methods which serves html files that are stored in the templates folder. The models folder stores the user class represented via SQLALchemy which is an object relational mapper between the python user object and the underlying table in the Postgresql database. Finally, the controller folder contain methods which alter the user model. A typical MVC scenario goes as follows: the client creates an account via the html render from the view methods, the request is sent to controller methods which in turn alter the user model class that finally updates the underlying table in Postgresql. This update in turn gets reflected in the view method as the user gets redirected to his or her newly created account page. Source code can be accessed at Github Link: https://github.com/LurenWang/ProjectReconnect.

3 Discussions

3.1 P-value and False Positive

For each of the 384 rounds of genotype matching between two individuals, only in two scenarios will a mismatch occur: when one individual is homozygous reference allele while the other is homozygous non-reference allele. P-value, the chances that two random individuals will survive all matches of 384 SNP position is 8.06E-21, calculated using eqn.1, with p_i being the reference allele frequency among the population.

$$P = \prod_{1}^{384} (1 - 2 \times p_i^2 \times (1 - p_i)^2) \quad (1)$$

Copy number variance accounts for 13% of human genome. In the worse case where both in parent and son 13% SNP failed to be probed, render a total 26% SNP genotyping failure or 284 successful calls, the combined probability of a random match between two individuals is still 2.69E-15; for a 7 billion world population, that corresponds to a 1.8e-5 False Positive.

3.2 Threshold establishment for a Positive Match

To give our platform some error tolerance capacity we loosen the criteria for a positive match from 100% to a lower value. From the minor allele frequency table, we can conclude that probability of a random successful match for a SNP position is equal or smaller than 0.897. We use 0.897, the upper bound probability to estimate the proper threshold for a positive match. A binomial distribution model with n = 384 and p = 0.897 was employed. Number of False Positive was plotted against the threshold of percentage of matched ones among the 384 SNPs.

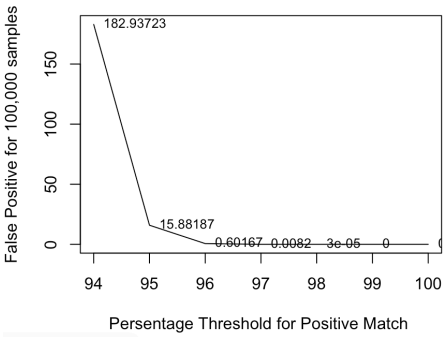

Fig. 1. Relation between percentage threshold and False Positive expectations for 100,000 samples . 98% was selected.

3.3 Other Kinship Scenarios

3.1.1 Siblings

We simulated the genotypes of 100,000 pairs of siblings and calculated their match percentage. The results can be viewed in Fig. 1 and table 2.

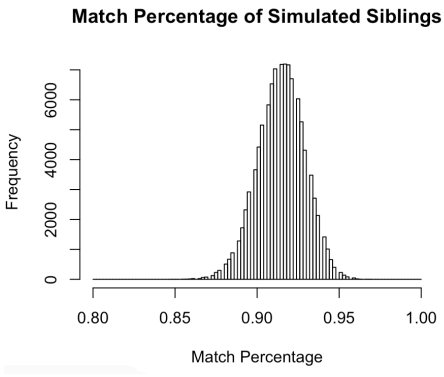

Fig. 1. match percentage of 100,000 simulated siblings

Table 2. Summary of match percentage of 100,000 simulated sibling pairs

Min.	1st Qu.	Mean	3rd Qu.	Max.
0.8464	0.9062	0.9146	0.9245	0.9740

3.2 Grandparents

We simulated the genotypes of 100,000 pairs of grandfather-grandson and calculated their match percentage. Results can be viewed in Fig. 2 and table 3.

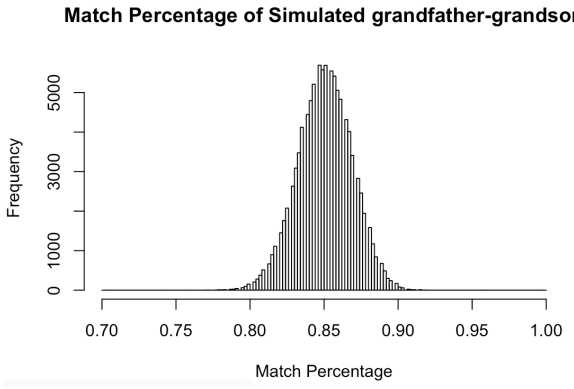

Fig. 2. match percentage of 100,000 simulated grandfather-grandson pairs

Table 3. Summary of match percentage of 100,000 simulated grandfather-grandson pairs

Min.	1st Qu.	Mean	3rd Qu.	Max.
0.7708	00.8385	0.8505	0.8620	0.9193

Match percentage obtained from the simulated samples indicate currently Project Reconnect threshold (98%) only captures parental relationship. We consider this result to be acceptable considering the core goal of Project Reconnect is to identify parents of lost Children.

Acknowledgements

Project Reconnect was completed under the advisement and with crucial support from Professor Itsik Pe'er at Columbia University. The authors would also like to acknowledge their classmates in the spring 2016 section of Professor Pe'er's Computational Genomics class and teaching assistant Shuo Yang for their advice and encouragement throughout this process.

References

Amigo, J., Salas, A. and Phillips, C. SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access. BMC … 2008.

Augustinus, D., Gahan, M.E. and McNevin, D. Development of a forensic identity SNP panel for Indonesia. International Journal of Legal Medicine 2015;129(4):681-691.

Boonyarit, H., et al. Development of a SNP set for human identification: A set with high powers of discrimination which yields high genetic information from naturally degraded DNA samples in the Thai population. Forensic Science International: Genetics 2014;11:166-173.

Bureau of Democracy, H.R.A.L. 2009 Human Rights Report: China (Includes Tibet, Hong Kong, And Macau). In.: United States Department of State 2009.

Consortium. A global reference for human genetic variation. Nature 2015.

Firth, H.V., et al. DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources. The American Journal of Human Genetics 2009;84(4):524-533.

Gibbs, R.A., et al. The international HapMap project. Nature 2003.

Huang, Q.Y., et al. Mutation patterns at dinucleotide microsatellite loci in humans. The American Journal of … 2002.

Hwa, H.-L., et al. Genotyping of 75 SNPs using arrays for individual identification in five population groups. International Journal of Legal Medicine 2016;130(1):81-89.

Jiang, Q. and Sanchez-Barricarte, J.J. Child Trafficking in China. China Report 2013;49(3):317-335.

Kim, J.-J., et al. Development of SNP-based human identification system. International journal of legal medicine 2010;124(2):125-131.

Kim, S. and Misra, A. SNP genotyping: technologies and biomedical applications. Annu. Rev. Biomed. Eng. 2007.

Liu, Z. 警方高压打拐人贩子被迫收手
. In, China Legal Daily. 2015. p. 1.

Machiela, M.J. and Chanock, S.J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics 2015.

Neumann, J. China's stolen children. In.; 2007.

Pakstis, A.J., et al. SNPs for a universal individual identification panel. Human genetics 2010;127(3):315-324.

Piñero, J., Queralt-Rosinach, N. and Bravo, À. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. … 2015.

Reich, D.E., et al. Human genome sequence variation and the influence of gene history, mutation and recombination. Nature … 2002.

Wei, Y.-L., et al. Validation of 58 autosomal individual identification SNPs in three Chinese populations. Croatian Medical Journal 2014;55(1):10-13.

Zeng, Z., et al. Evaluation of 96 SNPs in 14 Populations for Worldwide Individual Identification*. Journal of Forensic Sciences 2012;57(4):1031-1035.

Supplementary Table 1. Full List of 384 SNPs for paternal testing in Project Reconnect

rs.	Chromosome	rs.	Chromosome	rs.	Chromosome	rs.	Chromosome	rs.	Chromosome	rs.	Chromosome
rs4970357	1	rs12233824	4	rs2469364	8	rs519536	12	rs6500720	16	rs7260343	19
rs260513	1	rs10002315	4	rs1471593	8	rs12424642	12	rs3848386	16	rs6512121	19
rs2236395	1	rs4696803	4	rs12680074	8	rs4930768	12	rs17139223	16	rs11085743	19
rs2993480	1	rs940136	4	rs649126	9	rs2159953	12	rs1811321	16	rs6135141	20
rs2483256	1	rs6811287	4	rs7850218	9	rs10774398	12	rs1013614	16	rs723477	20
rs2493278	1	rs2672737	5	rs1412255	9	rs2072373	12	rs4786101	16	rs6085597	20
rs4323679	1	rs11746538	5	rs12344160	9	rs2305340	12	rs2192347	16	rs6134424	20
rs7524800	1	rs6554595	5	rs1331817	9	rs9511255	13	rs2131226	16	rs615007	20
rs11808641	1	rs12653484	5	rs10757469	9	rs6490551	13	rs17669927	16	rs214833	20
rs349393	1	rs11133883	5	rs3858032	9	rs7328274	13	rs4627375	16	rs6037397	20
rs780606	1	rs10076206	5	rs10974261	9	rs9552472	13	rs3951818	17	rs6139011	20
rs874668	1	rs10866539	5	rs10814874	9	rs1547149	13	rs2277669	17	rs6084327	20
rs2898853	1	rs1106115	5	rs301430	9	rs912134	13	rs9818	17	rs4815621	20
rs5026665	1	rs7712012	5	rs7848468	9	rs455692	13	rs1984749	17	rs9636534	20
rs1081454	1	rs11745728	5	rs295258	9	rs1579373	13	rs756819	17	rs2756271	20
rs3789553	1	rs13176914	5	rs1571221	9	rs943380	13	rs10491216	17	rs2102486	20
rs9434662	1	rs4866463	5	rs947403	10	rs4769245	13	rs8065080	17	rs805743	20
rs378930	1	rs11750441	5	rs2892336	10	rs2861537	13	rs12150491	17	rs6085352	20
rs4908611	1	rs261137	5	rs10903470	10	rs7317321	13	rs1050998	17	rs1342347	20
rs389709	1	rs1376240	5	rs11250681	10	rs2793483	13	rs3764900	17	rs6054479	20
rs4372938	2	rs4702249	5	rs12571115	10	rs7338915	13	rs9902174	17	rs6085681	20
rs11686452	2	rs6596835	6	rs3763685	10	rs12434613	14	rs887525	17	rs2775537	21
rs732609	2	rs10223528	6	rs7895038	10	rs4982401	14	rs17825449	17	rs2823035	21
rs2241457	2	rs9378518	6	rs17159312	10	rs10149170	14	rs6502926	17	rs2823252	21
rs12471985	2	rs10458234	6	rs10903959	10	rs7145515	14	rs1467138	17	rs2823795	21
rs2551212	2	rs2745631	6	rs1359526	10	rs17182867	14	rs6502958	17	rs2824110	21
rs10174217	2	rs9392312	6	rs1668538	10	rs3811313	14	rs2271314	17	rs2824242	21
rs1729926	2	rs9405506	6	rs12762783	10	rs2331662	14	rs2135845	17	rs4536738	21
rs7583267	2	rs12664420	6	rs1468064	10	rs17256106	14	rs222852	17	rs2150385	21
rs10197283	2	rs2296355	6	rs7916379	10	rs3825581	14	rs8066124	17	rs2824720	21
rs13425182	2	rs1040521	6	rs9423376	10	rs7142207	14	rs3027302	17	rs2825092	21
rs4371367	2	rs11242860	6	rs7917859	10	rs1033844	14	rs11654033	17	rs2825581	21
rs10172284	2	rs541051	6	rs7122936	11	rs8021991	14	rs4632172	17	rs2825731	21
rs1377638	2	rs1016598	6	rs4881743	11	rs206225	14	rs12945811	17	rs12482714	21
rs2882275	2	rs9378430	6	rs10769945	11	rs854351	14	rs7216753	17	rs1557330	21
rs941009	2	rs4959340	6	rs2651836	11	rs2208432	14	rs880334	17	rs2032196	21
rs751394	2	rs9504466	6	rs718579	11	rs17212261	14	rs8095469	18	rs1032002	21
rs332478	3	rs4637765	7	rs234853	11	rs178222	14	rs690196	18	rs175139	22
rs718559	3	rs10264122	7	rs12421922	11	rs17111632	14	rs1009238	18	rs5994128	22
rs6779648	3	rs2734791	7	rs11028536	11	rs7170864	15	rs8089407	18	rs5992916	22
rs17043656	3	rs6942930	7	rs1372804	11	rs487586	15	rs9955881	18	rs1557847	22
rs11129006	3	rs10260968	7	rs2278170	11	rs2316583	15	rs2282636	18	rs6518580	22
rs1948796	3	rs1982157	7	rs10768434	11	rs12595571	15	rs3786455	18	rs12167717	22
rs2616595	3	rs1550202	7	rs2472527	11	rs6576636	15	rs4798214	18	rs5759455	22
rs1391933	3	rs6952068	7	rs11037417	11	rs4514648	15	rs11877871	18	rs7289487	22
rs340813	3	rs17133257	7	rs317773	11	rs3097531	15	rs4321285	18	rs8141816	22
rs1386948	3	rs6976105	7	rs16933888	11	rs16950979	15	rs292292	18	rs131654	22
rs9862229	3	rs10272726	7	rs997433	11	rs11637374	15	rs1940994	18	rs4821718	22
rs2584043	3	rs12055961	7	rs1506981	11	rs8039737	15	rs9959692	18	rs8137866	22
rs12630241	3	rs6960751	7	rs523169	12	rs4780224	15	rs7504372	18	rs4822154	22
rs2600116	3	rs2042545	7	rs722097	12	rs2140175	15	rs4897966	19	rs5751614	22
rs6799552	3	rs9655516	7	rs10849549	12	rs4779527	15	rs4807121	19	rs2330555	22
rs6828710	4	rs9314442	8	rs7311263	12	rs8028396	15	rs20567	19	rs9612921	22
rs642626	4	rs7463086	8	rs4765829	12	rs7177176	15	rs1548709	19	rs507715	22
rs10805007	4	rs1470777	8	rs9805049	12	rs1317242	15	rs2873409	19		
rs2980098	4	rs7826946	8	rs2283275	12	rs2017567	16	rs6510698	19		
rs4689915	4	rs6558441	8	rs2238044	12	rs2235505	16	rs12461372	19		
rs3821949	4	rs10866924	8	rs215992	12	rs2729577	16	rs4807437	19		
rs16836929	4	rs3779704	8	rs11062386	12	rs2745170	16	rs3859564	19		
rs2269920	4	rs11780530	8	rs10774085	12	rs3094472	16	rs243342	19		
rs4689005	4	rs1455633	8	rs12810695	12	rs9928269	16	rs2620854	19		
rs2301795	4	rs9693133	8	rs503554	12	rs2191416	16	rs10403489	19		
rs4626203	4	rs12546362	8	rs758637	12	rs2238419	16	rs461970	19		
rs3892041	4	rs7013027	8	rs10849017	12	rs7186481	16	rs2860172	19		
rs7694661	4	rs11774749	8	rs10774223	12	rs3747590	16	rs7257153	19		
rs4689798	4	rs9314497	8	rs1468556	12	rs4786664	16	rs4804367	19		