A universal text-based identifier ("GenomeID") for human genomes

Paul Gordon (gordonp@ucalgary.ca) , preview v1

Impetus

As the number of whole genome and exome sequences grows exponentially, there is a need to be able to quickly determine if two datasets have been derived from the same individual. This may be for example because a cell line or participant has been used in more than one study, or simply to ensure no label swapping of biological samples. It is also possible that close relatives have not been identified as such in or across studies. In all of these situations, the inability to quickly identify a common source individual for datasets can lead to incorrect scientific inferences. This document defines how to calculate a short (12 or 20 characters) text string that can be used as an identifier for a whole genome DNA sample based solely on the information content of a genome or exome sequencing result.

Requirements

0. Intrinsic in Nature

Practical Requirement: The identifier must be computable solely from the genomic data, not relying on any other information outside of this specification for its calculation.

1. Individual Privacy

Practical Requirement: The identifier does not reveal the specific genetic makeup of the individual.

Formal Requirement: The individual genome to ID mapping function should be both non-injective and surjective. That is to say, there is a single computable ID for any individual, but it is not possible to unambiguously reconstruct the individual's genotypes from the ID.

2. ID Stability

Practical Requirement: The "genetic proximity" of two identifiers can be calculated, enabling matching of tumor and normal samples, and *optionally* the ability to impute close familial relations.

3. ID Generation

Practical Requirement: Incompletely genotyped individuals are able to have GenomeIDs.

4. ID Verification

Practical Requirement: Given partial knowledge of an individual's genome, the probability that it corresponds to a specific GenomeID can be calculated.

Proposed Solutions

Intrinsic: This document defines an identifier using only genotype information for the markers defined in the appendices.

Individual Privacy: The information stored in the identifier represents the zygosity of autosomal genotypes, not the genotypes themselves.

ID Stability: By using a collection of highly polymorphic, independent (low LOD score) and physically well spaced markers, identifier perturbations due to somatic changes are small compared to the difference between average individuals (see ID Statistics).

ID Generation: The identifier encodes information to indicate genetic markers that were not elucidated/available when the ID was generated.

ID Verification: A match using partial genome data can be easily verified probabilistically, based on a Bernoulli distribution.

Specification

A GenomeID is a 12 or 20 ASCII character string, encoded using Base64 (i.e. has allowed characters A-Z, a-z, 0-9, "+" and "/"). For example, the GenomeID of the widely used Coriell NA12878 cell line is:

ABL0N0yhmh9B
or
ABL0N0yhmh9BHmkolpuT

After Base64 decoding, the binary string consists of 72 or 120 bits in the following format:

Bits 1-6: *The missing autosomal data information block (MADIB)*. A six bit binary number, therefore in the range 0-63. The interpretation of the value is as follows:

- When 0, indicates that all markers bits have observed values (i.e. is complete).
- When in the range [1, 58], indicates the offset of the one marker bit that does not contain a valid zygosity observation. The value of the indicated marker bit must be zero.

- When 59, indicates that only the bits with defined data are those corresponding to the 24 markers defined in Pengelley et al. 2013. In this case, unused marker bits must have the value 0.
- When in the range [60, 62], indicates the number of missing markers by the calculation n - 58. This information can be used to adjust match probability statistics.
- When 63, indicates that there are more than 4 markers with missing data.

Bits 7-64: *Autosomal markers block (AMB)*. The zygosity information for 58 biallelic sites on the autosomal human chromosomes. Observed homozygous marker genotypes in an individual are indicated with a zero, heterozygous genotypes are indicated by a 1. The first marker occupies the leftmost bit in the AMB block, etc. The set of AMB marker positions is defined in Appendix A.

Bit 65: *Sex marker bit (SMB)*: 1 if Y chromosome is present ("male") [1], otherwise 0.

Bit 66: *Sex interpretation bit (SIB)*:

- For 12 character GenomeIDs, 0 if the sex of the individual is known (i.e. bit 65 is informative), otherwise 1. If 1 (sex unknown), bit 65 must be set to 0.
- For 20 character GenomeIDs, 1 if either sex chromosome has an unusual copy number that *affects normal genotype interpretation* of the XMB, otherwise 0. This is the only bit in the GenomeID that may not be derived directly and wholly from sequencing data alone.

Bits 67-72: *GenomeID Version Block (GVB).* A binary number representing the major (integer) version of the GenomeID specification with which the identifier is compatible.

Bits 73-120: *X markers block (XMB).* 24 biallelic X chromosome markers, with each marker encoded using two bits when SMB=0. The values of the two bits are to be interpreted as follows:

- 00: homozygous for ancestral allele
- 11: homozygous for alternate allele
- 01: heterozygous for both alleles
- 10: genotype undetermined

The 24 markers on chromosome X that are the most polymorphic in many subpopulations were taken from Tomas et al., 2010. The list of XMB marker positions is defined in Appendix B. The first marker occupies the first two leftmost bits in the block, etc.

When SMB=1, the high (left) bit from last 9 chrX markers instead contain bits that encode which Y SNP is present amongst the minimal Y haplotype tree published by van Oven et al. 2014. To decode, the last 9 XMB high bits are concatenated to form a binary integer, indicating the Y SNP index number. The indexed list of Y SNPs is

[1] Since the primary function of the sex markers is to infer family relationship

defined in Appendix C. The SNP index present with a deepest Y tree branch length should be used. If more than one SNP index number from *different* Y haplogroups are applicable to the sample, the first XMB high bit must be set to 1. All XMB high bits not described in this paragraph must be set to zero for samples from males. If a sample is known to be male but no existing marker is applicable (either due to lack of data, or a vary rare Y haplotype), then SMB=1, and all XMB high bits must be set to zero.

Notable observations

The first character of the identifier

Because of the nature of Base64 encoding, the first letter of the GenomeID contains only the MADIB, and hence indicates the completeness of the autosomal genotype data.

- GenomeIDs with full (58) marker information start with 'A'.
- GenomeIDs with only the 24 Pengelley marker observations start with '7'.
- GenomeIDs with two missing marker observations start with '8'.
- GenomeIDs with three missing marker observations start with '9'.
- GenomeIDs with four missing marker observations start with '+'.
- GenomeIDs with more than four missing marker observations start with '/'.
- GenomeIDs with a single missing marker will start with any other character.

The twelfth character of the identifier

This character encodes the sex of the individual. Karyotypically normal males will have a "B" for the twelfth letter, females an "h" if compatible with this version of the GenomeID specification. An "R" character in this position indicates that the sex of the individual is unknown. The example GenomeID used in this document (AuR08eFl1WWoh) is quickly identifiable as a full specified GenomeID for a female, based on the leading "A", and the trailing "h".

The thirteen character of 20 letter GenomeIDs

In karyotypically normal males with an unambiguous Y tree haplogroup, the thirteenth letter of the identifier will start with A-Z, or a-f. Note that this is NOT the Y haplogroup letter itself, rather the last nine XMB high bits must be decoded via Appendix C. Any other lower case letter, number or symbol in position 13 indicates that Y SNPs from multiple Y haplogroup tree branches were present in the genomic data (i.e. an unusual situation).

The last eight characters of 20 letter GenomeIDs

Because the nature of Base64 encoding, the last eight letters of the GenomeID contain only the sex chromosome marker data. This can be used to calculate the probability of familial relationship between any two individuals' GenomeIDs along maternal lineages, and paternal lineage for females. Because it contains genotype

data (unlike the AMB which only encodes zygosity), it can also be used to exclude parent-offspring relationships, based on presumed Mendelian inheritance. In other words, 12 character GenomeIDs have no intrinsically useful data content, whereas 20 character GenomeIDs have the potential to either preclude or suggest close relationships.

ID Statistics

The random probability of any two *specific* individuals[2] having the same 12 character GenomeID (i.e. only autosomal information) can be roughly[3] estimated as less than 1 in 2600 billion. The probability of *any* two individuals in a sample of size x having the same GenomeID can be calculated according to the same formula as for the Birthday Problem, e.g. for a sample size of 100,000 individuals:

$$1 - x! \times \frac{\binom{2.6 \times 10^{12}}{x}}{(2.6 \times 10^{12})^x} \text{ where } x = 100\,000$$

Which yields about a 0.2% chance of a GenomeID collision for a cohort of 100K individuals. For a cohort of 1 million, the chance of at least one collision rises to 17%.

When sex chromosome information is included in a GenomeID, the likelihood of random match reduced further. Using Tomas et al. (2010) calculations as a base, the 24 X-chromosome marker data alone will randomly coincide in about one in 10^6 males, and one in 10^9 females.

Parent-offspring relationships between two samples based on GenomeIDs can often be ruled out due to violation of Mendelian inheritance patterns.[4] In instances where a parent-offspring relationship *cannot* be ruled out, forensic genetics practice (Morling et al., 2002) is to describe the ratio of the presumed parent's probability of contributing the offspring genotype vs. the probability of a random individual contributing to it, a.k.a. the Paternity Index. Appendix D lists formulae to calculate Father-Daughter, Mother-Daughter and Mother-Son likelihood ratios for X chromosome markers.

Father-Son relationship can only possibly exist if the Y marker indexes match. Due to the highly variable allele frequency distribution of Y-markers in different subpopulations, no reliable Father-Son Paternity Index is calculable directly from GenomeIDs alone.

---

[2] Except monozygotic siblings and immediately consanguineous individuals

[3] Assuming 60% chance of individual marker zygosity match, and 58 markers $=0.6^{58}$

[4] Assuming no *de novo* mutations in the sex chromosome markers

Note that because certain Y haplogroup markers have highly skewed subpopulation frequencies, it is possible that a general ethnic or macro-geographic paternal bloodline origin for a sample may be inferred from its 20 character GenomeID.

References

Morling N, Allen R, Carracedo A, Geada H., Guidet F, Hallenberg C, Martin W, Mayr W, Olaisen B, Pascali V, Schneider PM. (2002) Paternity Testing Commission of the International Society of Forensic Genetics: recommendations on genetic investigations in paternity cases. *Forensic Sci. Int.* 129:148–157.

Pengelly RJ, Gibson J, Andreoletti G, Collins A, Mattocks CJ, Ennis S. A SNP profiling panel for sample tracking in whole-exome sequencing studies. Genome Med. 2013 Sep 27;5(9):89. doi: 10.1186/gm492

Tomas C, Sanchez JJ, Castro JA, Børsting C, Morling N. Forensic usefulness of a 25 X-chromosome single-nucleotide polymorphism marker set. Transfusion. 2010 Oct;50(10):2258-65.

van Oven M, Van Geystelen A, Kayser M, Decorte R, Larmuseau MH. 2014. Seeing the wood for the trees: a minimal reference phylogeny for the human Y chromosome. *Hum Mutat* 35(2):187-191. doi:10.1002/humu.22468

Appendix A - Autosomal Markers

Chr	Position (GRCh37)	(GRCh38)	rsID (dbSNP137)	Gene	Alleles	1K Genomes MAF
1	45973928	45508256	rs2275276	MMACHC*	A/G	0.49
1	67861520	67395837	rs2229546	IL12RB2	A/C	0.45
1	158582646	158612856	rs2251969	SPTA1	C/T	0.45
1	179520506	179551371	rs1410592	NPHS2	A/G	0.43
1	209968684	209795339	rs2013162	IRF6	A/C	0.4
1	228431095	228243394	rs1771455	OBSCN	C/T	0.37
2	49381585	49154446	rs1394205	FSHR	A/G	0.36
2	75115108	74887981	rs10194657	HK2	A/G	0.45
2	169789016	168932506	rs497692	ABCB11	A/G	0.48
2	215820013	214955289	rs10498027	ABCA12	A/G	0.33
2	227896976	227032260	rs10203363	COL4A4	C/T	0.45
3	4403767	4362083	rs2819561	SUMF1*	C/T	0.33
3	45989044	45947552	rs2234358	FYCO1	G/T	0.46
3	148727133	149009346	rs4938	GYG1*	A/G	0.31
4	5749904	5748177	rs4688963	EVC	A/G	0.46
4	86915848	85994695	rs10003909	ARHGAP24	C/T	0.32
5	13719022	13718913	rs30169	DNAH5	A/C	0.41
5	40981689	40981587	rs1061429	C7	A/C	0.35
5	55155402	55859574	rs1009639	IL31RA	C/T	0.39
5	82834630	83538811	rs309557	VCAN	A/G	0.48
5	138456815	139121126	rs3088052	SIL1*	A/G	0.43
5	171849471	172422467	rs17074773	SH3PXD2B	A/G	0.38
6	56471402	56606604	rs9382658	DST*	A/G	0.37
6	146755140	146434004	rs2942	GRM1	A/G	0.48
7	34009946	33970334	rs10265207	BMPER	C/T	0.43
7	48450157	48410560	rs17548783	ABCA13	C/T	0.48
7	100804140	101160859	rs1048303	AP1S1*	C/T	0.5
7	151254175	151557089	rs8961	PRKAG2*	C/T	0.42
8	94935937	93923709	rs4735258	PDP1	C/T	0.5
9	27202870	27202872	rs639225	TEK	C/T	0.44
9	77415284	74800368	rs7859201	TRPM6	A/C	0.45
9	100190780	97428498	rs1381532	TDRD7*	C/T	0.43
10	69926097	68166340	rs2673794	MYPN	A/G	0.5
10	100219314	98459557	rs10883099	HPSE2	A/G	0.46
11	6629665	6608435	rs1043388	ILK*	C/T	0.31
11	16133413	16111867	rs4617548	SOX6	A/G	0.47
11	30255185	30233638	rs6169	FSHB	C/T	0.4
12	993930	884764	rs7300444	WNK1*	C/T	0.44

12	52200742	51806958	rs60637	SCN8A	A/C	0.43
13	25466955	24892817	rs3742165	CENPJ	C/T	0.49
13	39433606	9532292	rs9532292	FREM2	A/G	0.4
14	50769717	50302999	rs2297995	L2HGDH*	A/G	0.43
14	76045858	75579515	rs2287016	FLVCR2	A/G	0.44
16	70303580	70269677	rs2070203	AARS*	C/T	0.46
17	7192091	7288772	rs222842	YBX2	C/T	0.5
17	42449789	44372421	rs5910	ITGA2B	C/T	0.4
17	71197748	73201609	rs1037256	COG1*	A/G	0.44
18	21413869	23833905	rs9962023	LAMA3	C/T	0.34
19	10267077	10156401	rs2228611	DNMT1*	A/G	0.46
19	33353464	32862558	rs11084673	SLC7A9	A/G	0.3
19	55441902	54930534	rs269950	NLRP7	C/T	0.43
20	6100088	6119441	rs10373	FERMT1	C/T	0.49
20	19970705	19990061	rs2076584	RIN2*	A/G	0.45
20	35865054	37236651	rs4608	RPN2	C/T	0.31
20	52786219	54169680	rs2296241	CYP24A1	A/G	0.47
21	44323590	42903480	rs4148973	NDUFV3*	G/T	0.48
22	21141300	20787012	rs4675	SERPIND1	C/T	0.45
22	37469591	37073551	rs4820268	TMPRSS6	A/G	0.46

* indicates ubiquituously expressed gene as designated by:

Ramsköld D, Wang ET, Burge CB, Sandberg R. (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol*. 5:e1000598.

This means that only 14/58 markers can be reliably expected in any given cell for mRNA-derived GenomeIDs. Collisions are much more likely in mRNA-derived GenomeIDs than as described in this specification which assumes full diploid genotypes.

Appendix B – X Chromosome Markers

rsID (dbSNP 137)	Position (GRCh37 chrX)	(GRCh38)	Ancestral Allele	Alternate Allele
rs2056688	4066743	4148702	G	A
rs2128519	5616964	5698923	G	A
rs1534285	11882557	11864438	T	C
rs763056	18151389	18133269	G	A
rs1373592	25779585	25761468	T	C
rs993010	34091679	34073562	T	C
rs1557054	35632777	35614660	C	T
rs1243792	40871567	41012314	G	A
rs925178	42848162	42988913	C	T
rs1207480	47686005	47826606	C	T
rs1936313	67200648	67980806	T	C
rs1977719	78397143	79141646	T	C
rs1372687	94465580	95210581	T	G
rs1857602	95114611	95859612	T	C
rs985425	104268001	105023319	T	C
rs933315	106598707	107355477	A	G
rs2190288	116521416	117387453	G	T
rs1991961	119826608	120692753	T	G
rs1931662	120864696	121730843	C	A
rs149910	126325138	127191155	G	T
rs1340718	139920048	140837883	T	C
rs1930674	144091597	145010077	G	C
rs1339597	145069663	145988145	G	A
rs1981452	151150133	151981661	G	T

Appendix C – Y Chromosome Haplogroup Markers

SNP index	rsID (dbSNP 137)	Position (GRCh37 chrY)	(GRCh38)	Ref> Alt	Y haplogroup	
1	rs371502065	14902417	12790484	C>T	A0	depth: 4
2	rs181335666	6788191	6920150	G>A	A0	depth: 2
3	rs181016083	18255814	16143934	C>T	A0	depth: 4
4	rs187287389	16292866	14180986	G>A	A0	depth: 2
5	rs369315948	21739754	19577868	G>C	A1	depth: 3
6	rs187409543	6846482	6978441	G>A	A1	depth: 3
7	rs72625368	2710154	2842113	G>A	A1'2'3'4	depth: 2
8	rs191505182	17947672	15835792	G>A	A1'2'3'4	depth: 2
9	rs2524861	4898665	4898965	C>G	A1'2'3'4	depth: 2
10	rs3905	21722268	19560382	T>C	A2	depth: 6
11	rs3897	18571026	16459146	T>C	A2	depth: 7
12	rs374971653	21881115	19719229	C>T	A2	depth: 8
13	rs189205028	6845936	6977895	T>C	A2	depth: 5
14	rs188292317	7589303	7721262	T>C	A2'3'4	depth: 3
15	rs3904	21722098		G>C	A3	depth: 7
16	rs2032619	21925500		T>C	A3	depth: 6
17	rs2032649	15029492		T>G	A3	depth: 7
18	rs369313031	21729839		T>G	A3	depth: 6
19	rs2032630	21866840		A>T	B,C,D,E,F	depth: 3
20	rs112236481	23497067		C>G	B	depth: 7
21	rs79947741	21932926		A>G	B	depth: 12
22	rs370186578	21774180		C>T	B	depth: 13
23	rs111725135	21767959		G>A	B	depth: 7
24	rs371646183	21869519		C>T	B	depth: 7
25	rs2032601	14869076		C>T	B	depth: 6
26	rs2032662	15014262		C>T	B	depth: 8
27	rs2032671	15437063		C>T	B	depth: 10
28	rs113717750	2657411		G>A	B	depth: 9
29	rs112887592	2711408		C>G	B	depth: 9
30	rs113134414	6982169		A>C	B	depth: 9
31	rs374589940	2844095		G>T	B	depth: 10
32	rs367912965	6753433		A>G	B	depth: 10
33	rs368217995	7606120		T>C	B	depth: 11
34	rs377707696	7769194		C>T	B	depth: 11
35	rs375423133	6790204		C>G	B	depth: 10
36	rs368790958	2736732		T>G	B	depth: 9
37	rs371100148	8156900		T>G	B	depth: 9
38	rs112194431	6768265		G>C	B	depth: 10

39	rs111691221	14850579	C>T	B	depth: 5
40	rs181987418	23120337	G>A	B	depth: 8
41	rs191044406	7599027	A>T	B	depth: 10
42	rs112298449	2798459	C>T	B	depth: 4
43	rs181258697	17201891	G>A	B	depth: 11
44	rs190560626	17421349	G>A	B	depth: 9
45	rs187723852	6859558	C>G	B	depth: 8
46	rs2032612	21866491	C>T	C	depth: 9
47	rs2032659	15576203	C>T	C	depth: 8
48	rs2032660	15575780	A>T	C	depth: 8
49	rs2032666	15437564	C>T	C	depth: 6
50	rs2032668	15437333	A>C	C	depth: 7
51	rs369611932	21742158	T>G	C	depth: 7
52	rs373681213	21749881	A>G	C	depth: 9
53	rs3899	7291534	G>T	C	depth: 9
54	rs182352067	6845955	G>A	C	depth: 9
55	rs2032595	14813991	C>T	C,D,E,F	depth: 5
56	rs4141886	14197867	G>A	C,F	depth: 6
57	rs2267802	2828425	A>G	D	depth: 7
58	rs2268591	24464597	G>C	D	depth: 11
59	rs201362608	21765281	T>A	D	depth: 9
60	rs2032614	21930287	T>C	D	depth: 10
61	rs2032621	21872738	T>C	D	depth: 8
62	rs3212291	14851526	C>G	D	depth: 9
63	rs9786118	8852858	C>T	E	depth: 13
64	rs9786371	14636457	C>T	E	depth: 12
65	rs2032617	21896261	G>T	E	depth: 9
66	rs13447370	21778998	C>T	E	depth: 7
67	rs9341316	22744939	T>G	E	depth: 15
68	rs368762706	21740450	A>C	E	depth: 9
69	rs375228668	21741703	G>C	E	depth: 12
70	rs374966967	21752644	G>C	E	depth: 10
71	rs377458712	6822948	C>T	E	depth: 14
72	rs2032639	21890177	G>A	E	depth: 8
73	rs2032640	21892572	C>T	E	depth: 15
74	rs9306841	21778998	C>G	E	depth: 7
75	rs16980577	21083420	T>A	E	depth: 8
76	rs16980473	14159846	C>T	E	depth: 9
77	rs16980621	14060308	A>C	E	depth: 10
78	rs16980588	16253694	G>A	E	depth: 14
79	rs16980589	16374426	C>T	E	depth: 17
80	rs16980502	17294958	C>T	E	depth: 15
81	rs16980406	21646058	T>A	E	depth: 16

82	rs368511174	6859957	T>C	E	depth: 15
83	rs371254614	6932821	G>C	E	depth: 17
84	rs768983	6818291	C>T	E	depth: 11
85	rs369197260	6932007	G>C	E	depth: 13
86	rs373648041	17664771	A>C	E	depth: 13
87	rs369637510	7729622	G>A	E	depth: 14
88	rs9786819	14197977	G>A	E,N	depth: 17
89	rs2032652	21917313	C>T	F	depth: 7
90	rs56221698	6345148	C>G	G	depth: 13
91	rs2032636	15027529	G>T	G	depth: 9
92	rs13447378	22741740	G>C	G	depth: 10
93	rs368036847	23243942	C>T	G	depth: 10
94	rs368682405	15027433	A>G	G	depth: 11
95	rs4116820	22072097	G>T	G	depth: 10
96	rs9785956	15058878	A>G	G	depth: 16
97	rs2178500	23973594	T>G	G	depth: 11
98	rs376563811	23496331	C>T	H	depth: 13
99	rs13447371	21764431	A>C	H	depth: 11
100	rs376769460	21753199	A>C	H	depth: 12
101	rs35960273	14640715	A>G	I	depth: 13
102	rs2032597	14847792	A>C	I	depth: 12
103	rs367573274	21717307	G>A	I	depth: 16
104	rs9341274	15591446	C>G	I	depth: 16
105	rs9341296	15022707	C>T	I	depth: 13
106	rs9341301	15023364	T>C	I	depth: 12
107	rs2032629	21865821	G>A	I	depth: 16
108	rs373764666	15426005	C>T	I	depth: 17
109	rs2319818	16354708	G>A	I	depth: 12
110	rs17250163	21225770	C>G	I,J	depth: 11
111	rs377565422	5359116	C>T	J	depth: 14
112	rs372746826	5274763	G>A	J	depth: 15
113	rs199525748	6932171	G>C	J	depth: 17
114	rs368294560	15541333	C>T	J	depth: 14
115	rs369778229	9879005	C>T	J	depth: 19
116	rs2032608	21926112	G>C	J	depth: 14
117	rs3903	7583480	G>T	J	depth: 14
118	rs375691065	21716366	G>A	J	depth: 18
119	rs2032657	15587509	T>A	J	depth: 15
120	rs8179022	15018459	G>A	J	depth: 15
121	rs9341313	22741818	T>G	J	depth: 13
122	rs13447373	15467785	T>A	J	depth: 18
123	rs371079691	2751678	A>G	J	depth: 14
124	rs371666197	17419934	T>C	J	depth: 13

125	rs2032648	21904023	T>C	J	depth: 19
126	rs17315835	19179335	T>C	J	depth: 12
127	rs374310420	4065584	G>A	J	depth: 16
128	rs35192147	4929023	G>A	J	depth: 18
129	rs367604111	16359006	G>T	J	depth: 17
130	rs373832646	22797639	C>T	J	depth: 19
131	rs369454960	14555305	C>T	J	depth: 19
132	rs3900	21730257	C>G	K,L,T	depth: 11
133	rs3913	14790163	A>G	L	depth: 1
134	rs376303746	21739646	C>G	L	depth: 17
135	rs377156966	2888252	C>A	L	depth: 17
136	rs377001539	21757316	T>G	L	depth: 17
137	rs2032611	21866424	A>G	M	depth: 16
138	rs377239181	17309267	G>A	M	depth: 16
139	rs3895	2744628	A>G	M	depth: 16
140	rs13304043	9004921	G>A	M,S	depth: 16
141	rs371086975	19431608	A>C	N	depth: 17
142	rs377214926	21741755	C>T	N	depth: 19
143	rs2032674	15471925	T>C	N,O	depth: 15
144	rs2075181	7546726	C>T	O	depth: 20
145	rs373112617	21907097	T>C	O	depth: 19
146	rs72613040	21762685	A>C	O	depth: 18
147	rs78149062	21764674	T>C	O	depth: 17
148	rs2032605	14924869	C>T	O	depth: 20
149	rs13447443	22739301	A>G	O	depth: 18
150	rs13447361	2821786	G>C	O	depth: 18
151	rs2032632	21868672	T>C	O	depth: 19
152	rs3898	4078217	C>G	O	depth: 21
153	rs2032645	21900849	A>G	O	depth: 22
154	rs2032650	21938444	C>T	O	depth: 21
155	rs3212283	21889536	G>C	O	depth: 21
156	rs17316007	14001232	T>C	O	depth: 20
157	rs16981290	7568568	C>A	O	depth: 16
158	rs17276358	19179463	G>T	O	depth: 18
159	rs17316592	18560005	T>G	O	depth: 18
160	rs200861659	14495243	T>C	O	depth: 18
161	rs34893929	15999244	G>A	O	depth: 22
162	rs367562925	21285962	A>T	O	depth: 20
163	rs35724598	7014317	G>C	Q	depth: 18
164	rs34954951	23292782	G>A	Q	depth: 21
165	rs34864948	15574102	G>A	Q	depth: 19
166	rs374472129	21907394	T>C	Q	depth: 21
167	rs376134474	21889818	G>T	Q	depth: 20

168	rs3910	21733231	T>A	Q	depth: 24
169	rs8179021	15018582	C>T	Q	depth: 16
170	rs371369701	21866664	G>C	Q	depth: 20
171	rs3894	19096363	G>A	Q	depth: 23
172	rs4252209	4925637	G>T	Q	depth: 18
173	rs2032631	21867787	G>A	Q,R	depth: 15
174	rs368558585	4929885	G>A	R	depth: 19
175	rs13304168	14641193	C>T	R	depth: 24
176	rs9786111	19054889	C>T	R	depth: 22
177	rs372706460	21764501	C>T	R	depth: 18
178	rs2020857	15030752	C>T	R	depth: 20
179	rs9786153	22739367	T>C	R	depth: 21
180	rs9786184	2887824	C>A	R	depth: 18
181	rs17316771	8533735	G>A	R	depth: 21
182	rs9786486	4446430	T>A	R	depth: 22
183	rs9785702	18656508	G>C	R	depth: 20
184	rs185037316	18099054	C>T	R	depth: 20
185	rs79683764	14181107	C>T	R	depth: 27
186	rs9341277	15481372	T>A	S	depth: 16
187	rs369616152	14001024	T>A	S	depth: 16
188	rs374630929	6740377	T>C	S	depth: 16
189	rs2215828	19372808	C>T	T	depth: 16
190	rs34179999	16019072	G>C	T	depth: 15
191	rs35815655	6736443	C>T	T	depth: 16
192	rs2032672	21893881	A>C	T	depth: 14

Appendix D – Paternity/Maternity Index (likelihood ratio) Formulae Using X Chromosome Markers

Note: Father-Son and Mother-Son relationships involve non-recombinant sex chromosome inheritance, therefore likelihood ratios do not apply. Mendelian inheritance pattern violations should be used as parent-offspring exclusion criteria in those cases.

Father-Daughter

Mother-Daughter