# SUPPLEMENTARY METHODS AND ANALYSIS

**Reference Sequence:**

In absence of an explicit reference sequence declared by the 1KG-P3 project, we built our own reference. Using the UCSC gene and gene prediction table, we identified start and end points of all identified coding regions in the genome. Multiple overlapping/redundant annotations were merged to create a "meta" track of start and end points, which were extracted from the NCBI Human Genome Reference \ref{} to construct a new 123.7 Mbp reference sequence. Reads from each individual were subsequently aligned against this reference using Maq \ref{Maq} (refer Supplementary Data). The purpose of shortlisting a candidate reference sequence in this manner instead of using the whole genome was to minimize ambiguous read-alignment due to repeats.

**Datasets:**

We downloaded one lane of Solexa 51bp single-end short read datasets for each of 12 individuals from the SRA, as available in Jan 2009. The dataset accession numbers were: SRR003504, SRR003506, SRR003509, SRR003512, SRR003515, SRR003518, SRR003521, SRR003524, SRR003527, SRR003530, SRR003533 and SRR003536. Each individual has been sequenced 3 times, out of which reads from one run chosen at random were downloaded. The number of reads varies from 6.3M to 14.4M reads per run. Of these, the number of reads that aligned to our reference sequence using maq with default alignment parameters (upto 2 mismatches per read, default read filtering based on base quality) varies from XXX to YYY \ref{table}. For each individual we shortlisted sites which enjoyed more than 3X coverage as high confidence coding regions sequenced by the 1KG-P3 (approximately 5Mbp for each individual. Combining sites with >=3X coverage on atleast one (possibly multiple) individuals gives us 6.41x10^6 sites. This concurs with what we expect to be the cumulative size of the coding region of 1000-2000 genes, as stated by the 1KG-P3 project outline.

| Individual | No. of Reads | Mapped Reads | Sites with >=3X covg. |
|---|---|---|---|
| SRR003504 | 6350948 | 2065477 | 4445432 |
| SRR003506 | 6599672 | 2132322 | 4461256 |
| SRR003509 | 8923040 | 2851205 | 4805085 |
| SRR003512 | 10947166 | 3720192 | 5213638 |
| SRR003515 | 9002926 | 2022622 | 4531866 |
| SRR003518 | 10606950 | 3452037 | 5269475 |
| SRR003521 | 7398092 | 2277639 | 4573201 |
| SRR003524 | 6411922 | 1956323 | 4448326 |
| SRR003527 | 13223281 | 1911968 | 4497189 |
| SRR003530 | 14477735 | 899369 | 3477187 |
| SRR003533 | 11960197 | 3598033 | 5168564 |
| SRR003536 | 11598299 | 3667314 | 5161453 |

**Pool Designs:**

Given a12-individual dataset, we designed 2 sequencing arrangements of 8 pools, as it provided for both Logarithmic (unique column vectors) and Error Correcting assignments (unique column vectors, all of equal magnitude, with any 2 vectors being a minimum predefined distance apart) and an unbiased comparison of the two approaches.

The (Pools × Individuals) Design Matrices for each are shown below:

| Pool | 3504 | 3506 | 3509 | 3512 | 3515 | 3518 | 3521 | 3524 | 3527 | 3530 | 3533 | 3536 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Log1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| Log2 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| Log3 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| Log4 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Log5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| Log6 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| Log7 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| Log8 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |

Fig**: Logarithmic Pool Design**. Individuals are labeled by their accession numbers, without the prefix (e.g. "SRR003504" is "3504".) Each individual is sequenced on 4 pools. We note that the number of individuals sequenced in a pool varies from 4 (pools Log1 and Log2) to 8(pools Log5 and Log6). Correspondingly, unequal coverage is assigned to individuals depending on which pools they are sequenced in.

| Pool | 3504 | 3506 | 3509 | 3512 | 3515 | 3518 | 3521 | 3524 | 3527 | 3530 | 3533 | 3536 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ECC1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| ECC2 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| ECC3 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |
| ECC4 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| ECC5 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| ECC6 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| ECC7 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| ECC8 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |

Fig: **Error Correcting Pool Design.** Each individual is sequenced on 6 pools. Here, we note that the number of individuals sequenced in a pool is *also* constant across all pools. Correspondingly, equal coverage (subject to PCR and undersampling noise) is assigned to individuals regardless of which pools they are sequenced in.

Throughput per lane was kept constant, and varied from 6.3M to 14.4M reads per lane for the 12 individuals, and between 8M and 12M reads per lane for Log and ECC pools.

The results of the two designs were compared against the **Identity Design**, a 12 × 12 Identity matrix. This constitutes the dataset, where each sequencing lane has reads from just 1 individual.

**Pooled SNP calling:**

We built our own SNP caller for Pooled short-read data, borrowing from familiar concepts of SNP calling, and also introducing some new ones. SNPs on each pool in a design were called independently of the other pools in the design. To maintain sanctity of the experiment, the same SNP caller was used to ascertain SNPs on the Identity Pools (1 individual, 1 pool) as well as the Logarithmic and ECC design pools (multiple individuals, 1 pool).

For each pool, given (a) number of contributing individuals and (b) mean coverage across all short listed sites (≥3X coverage) in the alignment, our algorithm first filters sites based on observed coverage features (overall coverage, coverage per chromosome, allele coverage on forward and backward strands). We normalize under the assumption that sites of lower than expected coverage are undersampled on some chromosomes.

1. Multiple levels of noise filters then use entropy thresholds on allelic calls made at each site based on profiles of forward, reverse and both strands combined. We remove calls due to bad mapping, particularly where the calls are more than bi-allelic (e.g. 6 covering reads call AAACCG against an expected T on the reference is filtered out) while retaining sites with occasional errors due to bad reads (e.g. AAAAAC against a reference sequence T is retained, on the assumption that one read had a sequencing error.)
2. We then estimate the number of alleles at the site by using a maximum log likelihood ratio estimate. Consider a site at which C mapped reads call an allele w.r.t. the reference. Given a normalized estimate of N chromosomes covering a site, we establish N prior probabilities $p_x=x/N$ , $x \in \{1 \ldots N\}$, that x of these chromosomes carry the variant. Given a very conservative sequencing error prior probability of $\varepsilon=0.01$ (i.e. 1% sequencing error) per site \ref{Solexa Documentation}, we then calculate the LLR that x chromosomes carry the variant as

   LLR of x variant carrying chromosomes = $\text{Binom}(C, p_x)/\text{Binom}(C, \varepsilon)$

We estimate the number of variant carrying chromosomes as the one with most likely LLR score. Negative scores are indicative of sequencing errors rather than true positive allele carrier. The intricacies of our pooled SNP calling algorithm will be discussed in a future work. The source code and exectuable scripts for the algorithm are available on http://ron.cs.columbia.edu/papers/supplementary/

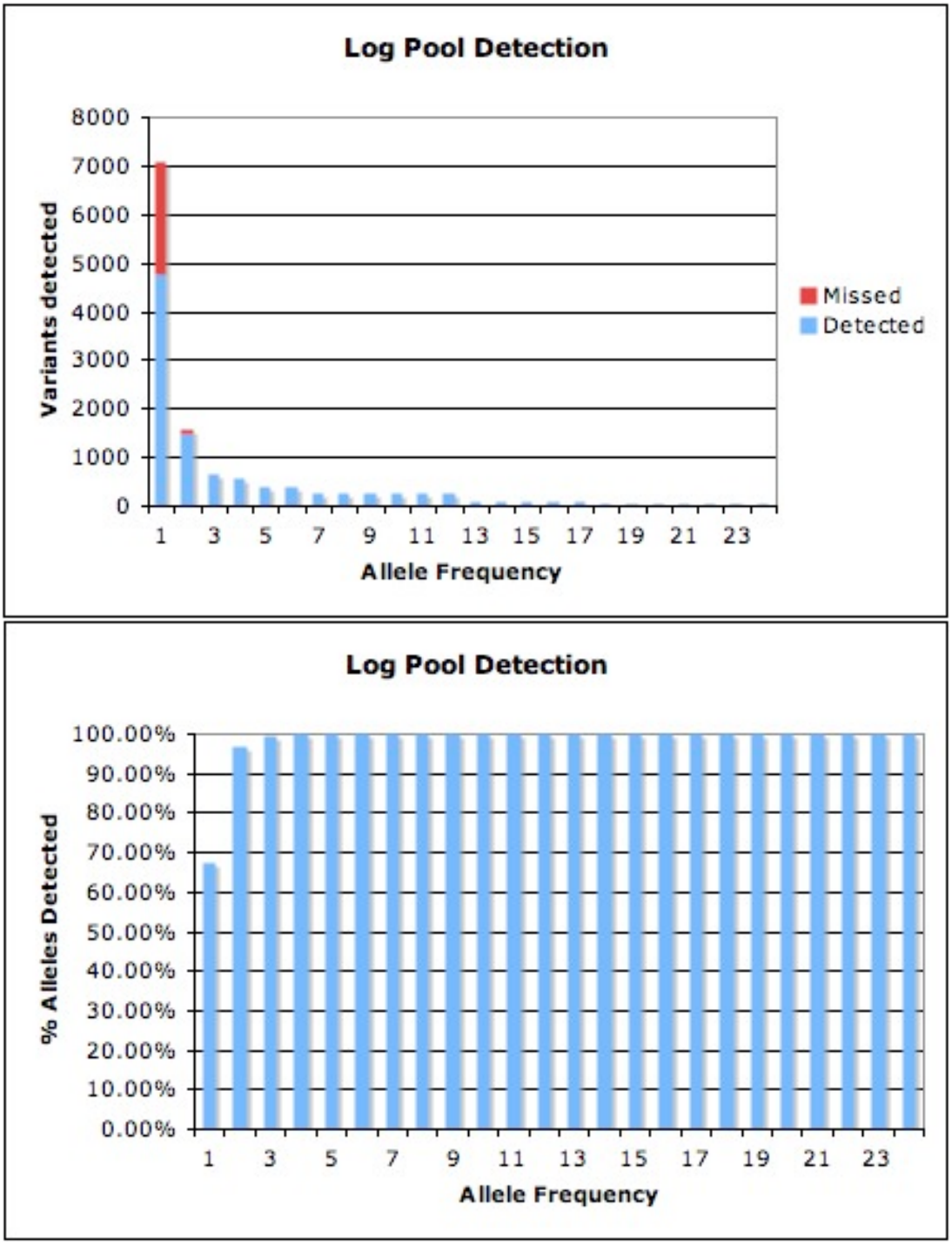
**Allele Detection:**

We first assessed the ability of the designs to detect the presence of a variant in any of the pools, regardless of how common the variation or who the variant carrier is. We called a total of 13022 SNPs (of varying confidence) across all 12 individuals, using the Identity Pools (parent dataset). The Log Designs detected a total of 10668 of these. Analysis of the undetected (false negative) variants shows that they are mostly rare (singleton) and

low confidence SNP calls, suffering from low coverage on forward, backward or both strands. Error-Correcting Design reported similar variant detection figures, with 10868 detected variants, and mostly the same SNPs going undetected. This demonstrated that certain profiles of low coverage SNPs do get missed.

The following table gives a summary of the ability of the Designs to detect variation.

| **Allele Freq.(in chromosomes)** | **Identity (actual)** | **Log (true positive)** | **Log recovery(%)** | **ECC (true positive)** | **ECC recovery(%)** |
|---|---|---|---|---|---|
| 1 (singletons) | 7062 | 4762 | 67.4% | 4966 | 70.3% |
| 2 (sometimes single homozygous carrier) | 1556 | 1508 | 96.9% | 1512 | 97.2% |
| 3 | 660 | 655 | 99.2% | 653 | 98.9% |
| 4 | 589 | 588 | 99.8% | 585 | 99.3% |
| 5 | 410 | 410 | 100.0% | 408 | 99.5% |
| 6 | 376 | 376 | 100.0% | 375 | 99.7% |
| 7 | 262 | 262 | 100.0% | 262 | 100.0% |
| 8 | 269 | 269 | 100.0% | 269 | 100.0% |
| 9 | 262 | 262 | 100.0% | 262 | 100.0% |
| 10 | 255 | 255 | 100.0% | 255 | 100.0% |
| 11 | 258 | 258 | 100.0% | 258 | 100.0% |
| 12 | 262 | 262 | 100.0% | 262 | 100.0% |
| 13 | 80 | 80 | 100.0% | 80 | 100.0% |
| 14 | 104 | 104 | 100.0% | 104 | 100.0% |
| 15 | 78 | 78 | 100.0% | 78 | 100.0% |
| 16 | 87 | 87 | 100.0% | 87 | 100.0% |
| 17 | 68 | 68 | 100.0% | 68 | 100.0% |
| 18 | 64 | 64 | 100.0% | 64 | 100.0% |
| 19 | 57 | 57 | 100.0% | 57 | 100.0% |
| 20 | 61 | 61 | 100.0% | 61 | 100.0% |
| 21 | 43 | 43 | 100.0% | 43 | 100.0% |
| 22 | 59 | 59 | 100.0% | 59 | 100.0% |
| 23 | 43 | 43 | 100.0% | 43 | 100.0% |
| 24 | 57 | 57 | 100.0% | 57 | 100.0% |




Figs: **Detection in Log Pools.** Both axes are marked in denominations of number of variant carrying chromosomes (i.e. 0 to 24 for 12 individuals). Each individual occurs in 4 pools. The black lines demonstrate 1 standard deviation above and 1 standard deviation below the mean prediction.

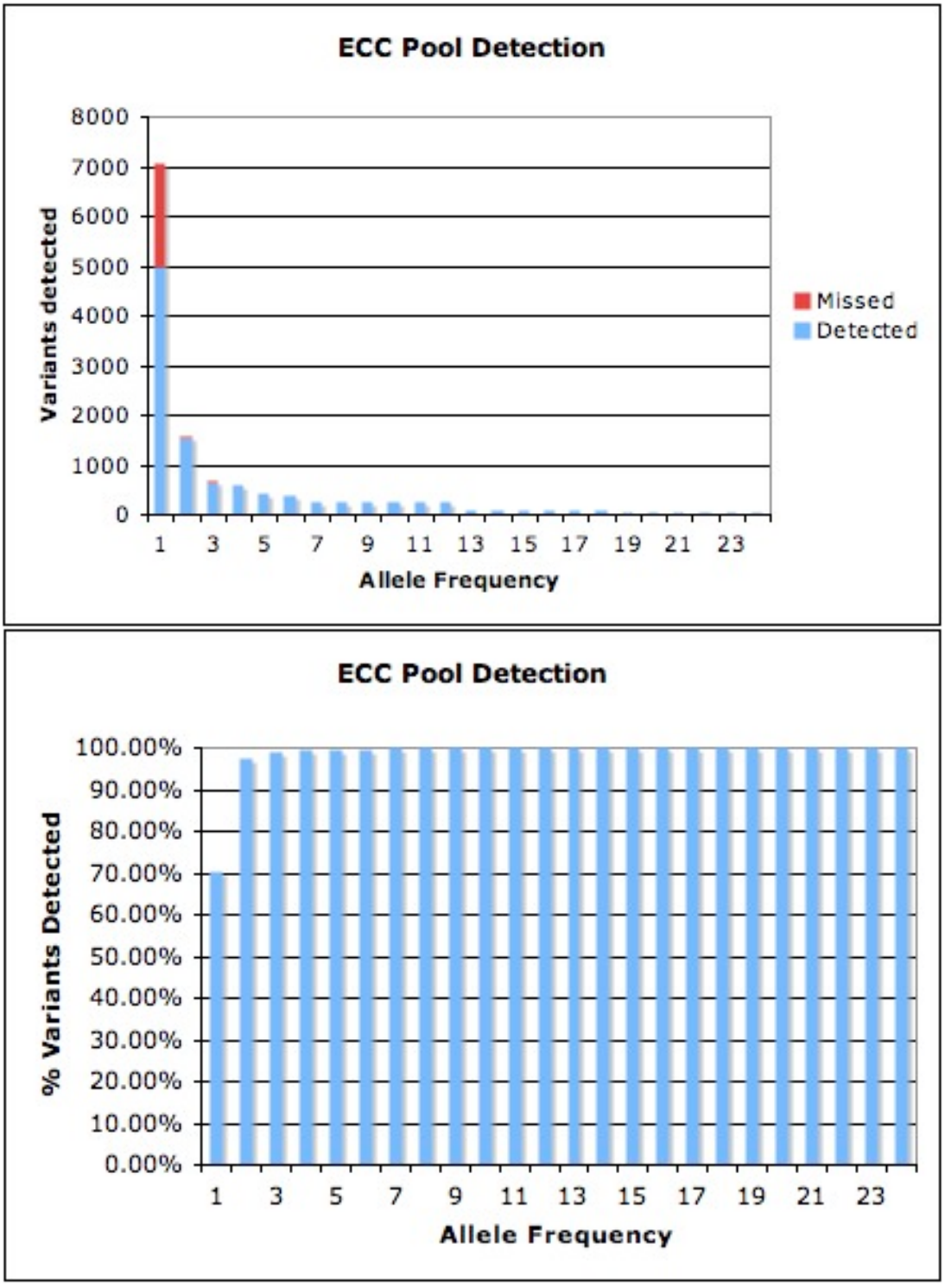



Fig: **Allele Frequency Estimate of Log Pools.** Both axes are marked in denominations of number of variant carrying chromosomes (i.e. 0 to 24 for 12 individuals). Each individual occurs in 4 pools. The black lines demonstrate 1 standard deviation above and 1 standard deviation below the mean prediction.

**Allele Frequency Determination:**

We then assessed the ability of our designs to predict the frequency of the occurring variations. The true frequency of an allelic site was calculated by summing the incidence of the total variants on the 12 single lane datasets. Summing the total number of variants at each pool and dividing by the number of pools per individual calculated the allele frequency estimate of a design. The allele frequency results demonstrate that both the pool designs are able to predict allele frequency with very good precision.

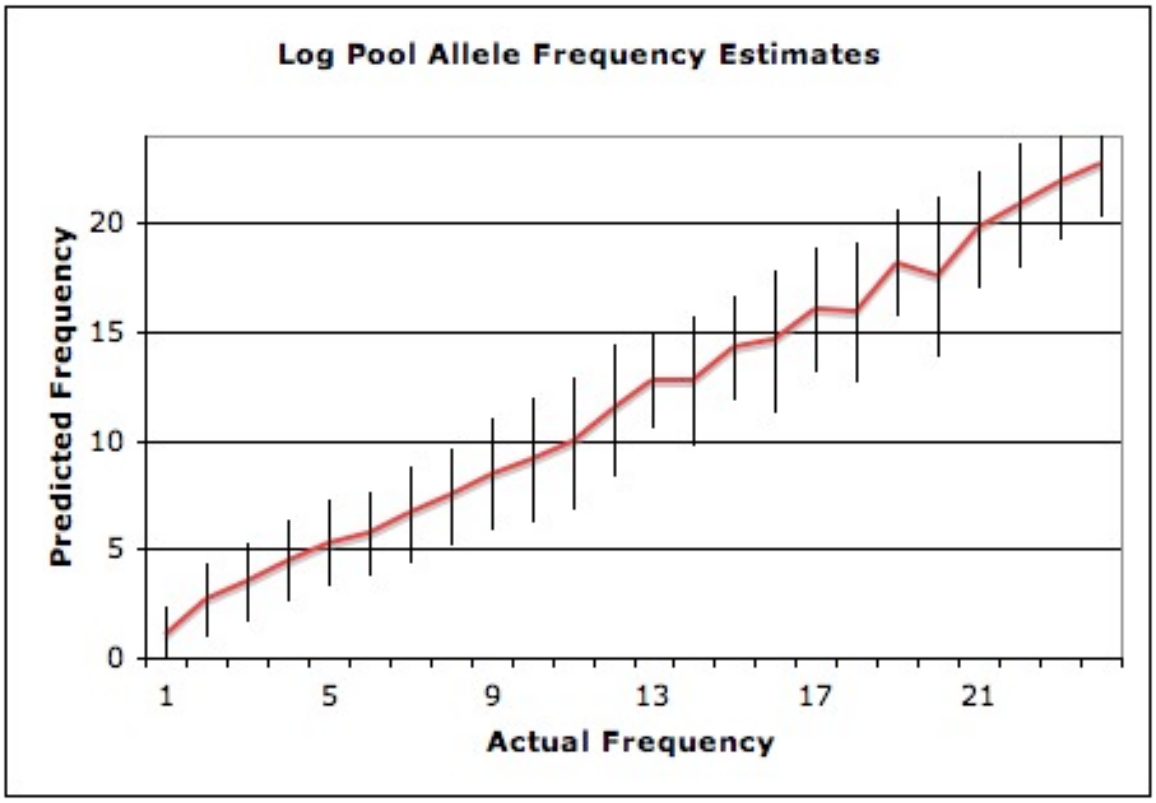



Fig: **Allele Frequency Estimate of Log Pools.** Both axes are marked in denominations of number of variant carrying chromosomes (i.e. 0 to 24 for 12 individuals). Each individual occurs in 4 pools. The black lines demonstrate 1 standard deviation above and 1 standard deviation below the mean prediction.

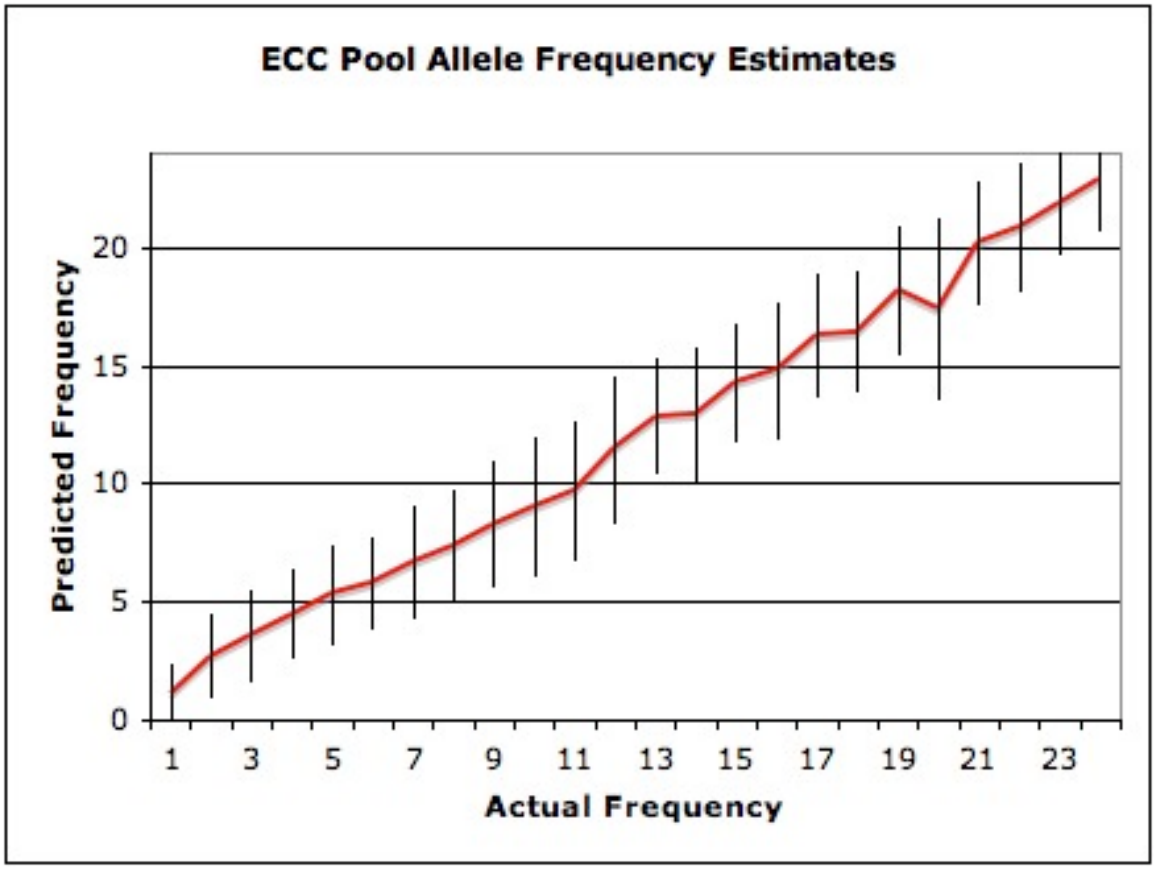



Fig: **Allele Frequency Estimate of ECC Pools.** Both axes are marked in denominations of number of variant carrying chromosomes (i.e. 0 to 24 for 12 individuals). Each individual occurs in 6 pools. The black lines demonstrate 1 standard deviation above and 1 standard deviation below the mean prediction.

**False Positive Calls:**

False positives were a major concern that the SNP caller had to deal with. It is a non-trivial problem to call sites that are sparsely covered or have few allele calling reads in pooled data. In the case of Identity pools (i.e. the dataset), it is easy to discard sites as false positives since we expect approximately half the covering reads (in the case of a heterozygous allele carrier) or all the covering reads (homozygous allele carrier) to call the variant. However, when there are multiple individuals contributing to an alignment, careful pruning of sites is called for.

While filtering for stringent coverage requirements are a certain way to ensure low False Positive rate, the tradeoff is incurring a high False Negative rate. Permitting a large numbers of False positive calls at the first stage through loose use of filtering captures most of the rare variation as well. We observed that a disproportionately large number of False positives called by the pools were predicted as singletons (77% of False Positives for Log, 72% for ECC) or doubletons (15% of False Positives for Log, 18% for ECC) in

frequency, calculated out of 24 chromosomes. This is because typically low coverage/rare variants in an alignment are falsely called as alleles.

Leveraging this observation, our pipeline loosely filters out sites at the initial stage, permitting large quantities of false calls. Since a vast majority of these calls are predicted singletons/doubletons, our pipeline then discards sites at the next stage, when we are unable to match them to an individual using the carrier detection algorithm discussed in the next section. This strategy permits us to reduce false positive rate by several orders of magnitude.

Analysis of False positive singletons shows that they are sites where alleles are called with disproportionately low coverage in the 12 downloaded datasets compared to True positive singleton calls.

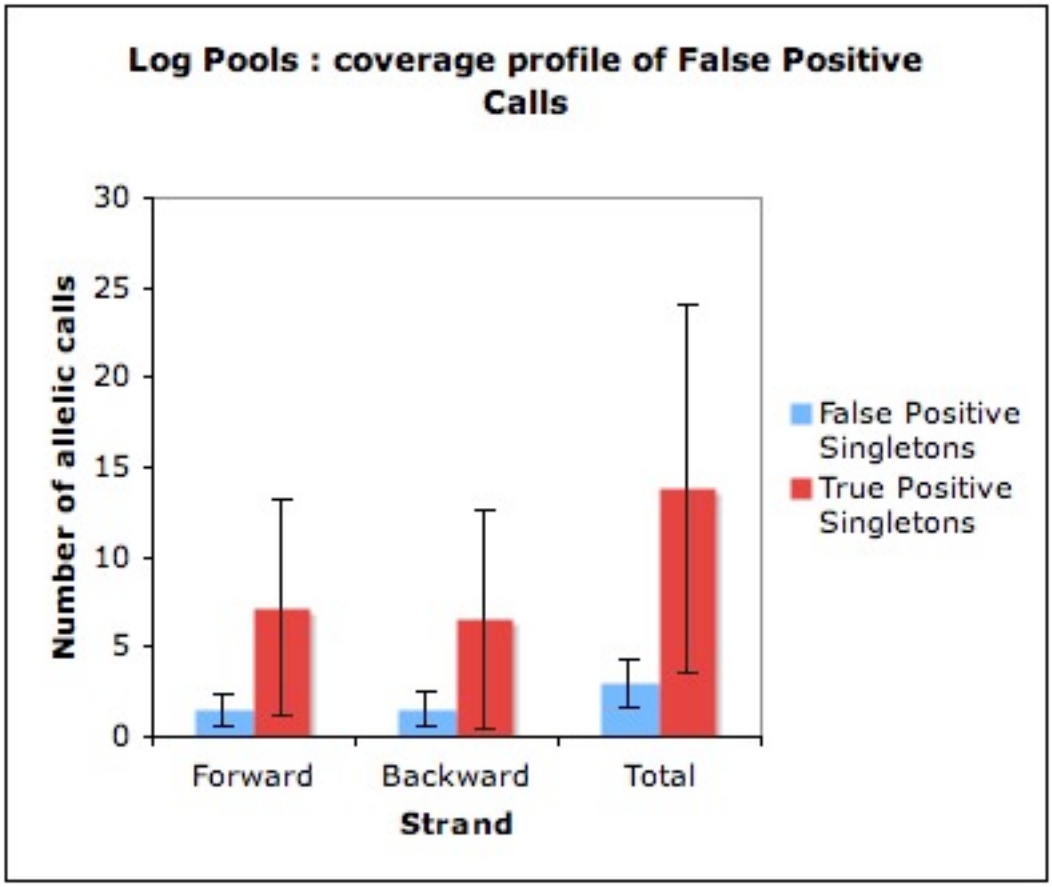



Fig: **False Positive Singletons on Log Pools.** Number of allele calling reads on the downloaded datasets of True positive singletons on both strands is much higher than on sites falsely called as singletons. The black bars show one standard deviation above and below the mean.

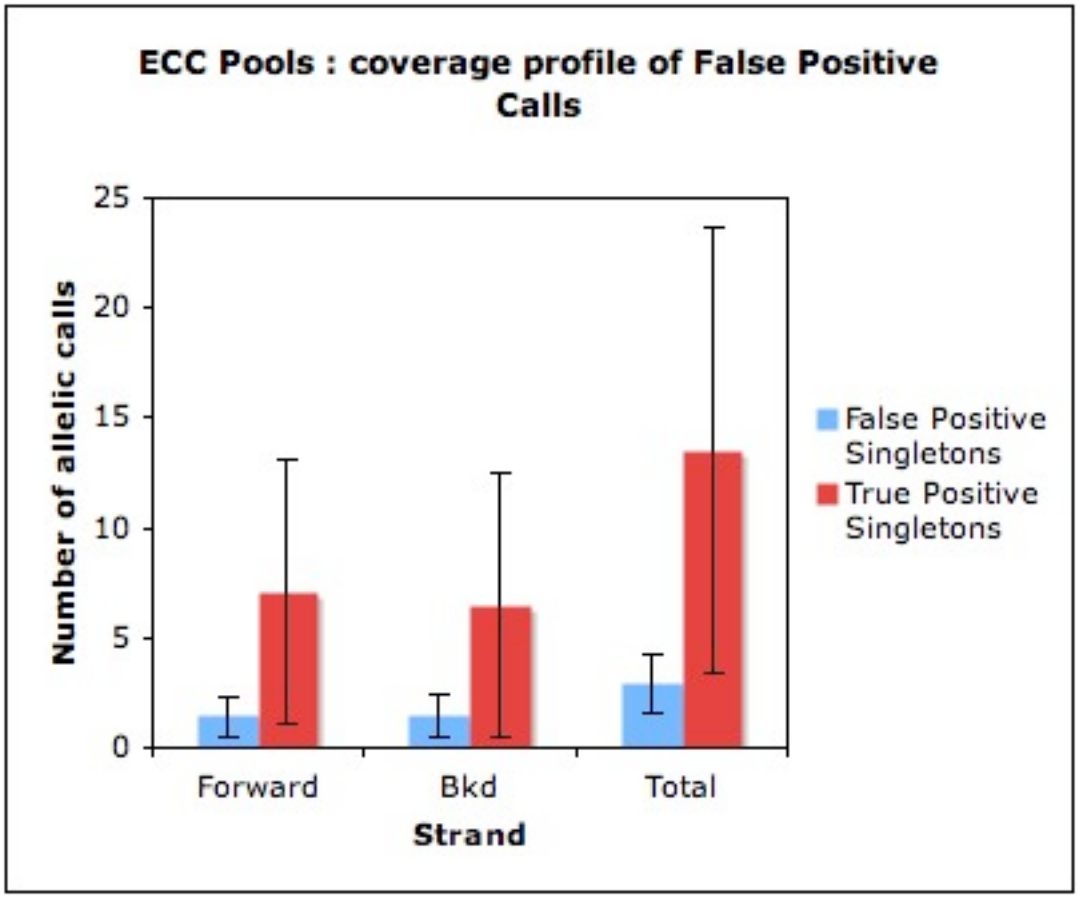



Fig: **False Positive Singletons on ECC Pools.** Number of allelic reads covering of True positive singletons in the datasets is higher on both strands than on sites falsely called as singletons. The black bars show one standard deviation above and below the mean.

**Carrier Identification:**

Based on the pool signature of each detected variant, we associated a distribution over possible carrier individuals. Out of a total of 8618 singleton and doubleton variants, Log Pools detected 6270 variants, while ECC Pools detected 6478 variants (refer table in section on Allele Detection). In truth (using Identity dataset), we ascertained that 5332 of the variants detected by Log pool had a single carrier (either homozygous causing singleton or heterozygous causing doubleton), while 5539 of the variants detected by ECC Pools had a single carrier individual.

At each of these sites, our algorithm uses the variant's pool signature to outputs a set of equally likely candidate individuals (uniform distribution) to be the variant carriers. Log Pools associated 4798 variants with a candidate carrier distribution while being unable to assign the rest. Likewise, ECC pools assigned 5060 variants with a distribution. In some cases the call is ambiguous (multiple individuals are given a uniform probability of being potential carriers), while in other case, the design identifies a single variant carrier.

Out of these calls, 3130 distributions in Log design captured the correct individual as 1 of the prospective carriers, while 2907 distributions in ECC design captured the same. Some variants strongly identified single individuals as their carriers, instead of offering a distribution over multiple prospective individuals. The fidelity of these calls show a strong correlation to what coverage the site enjoyed on the carrier individual dataset, before being pooled.

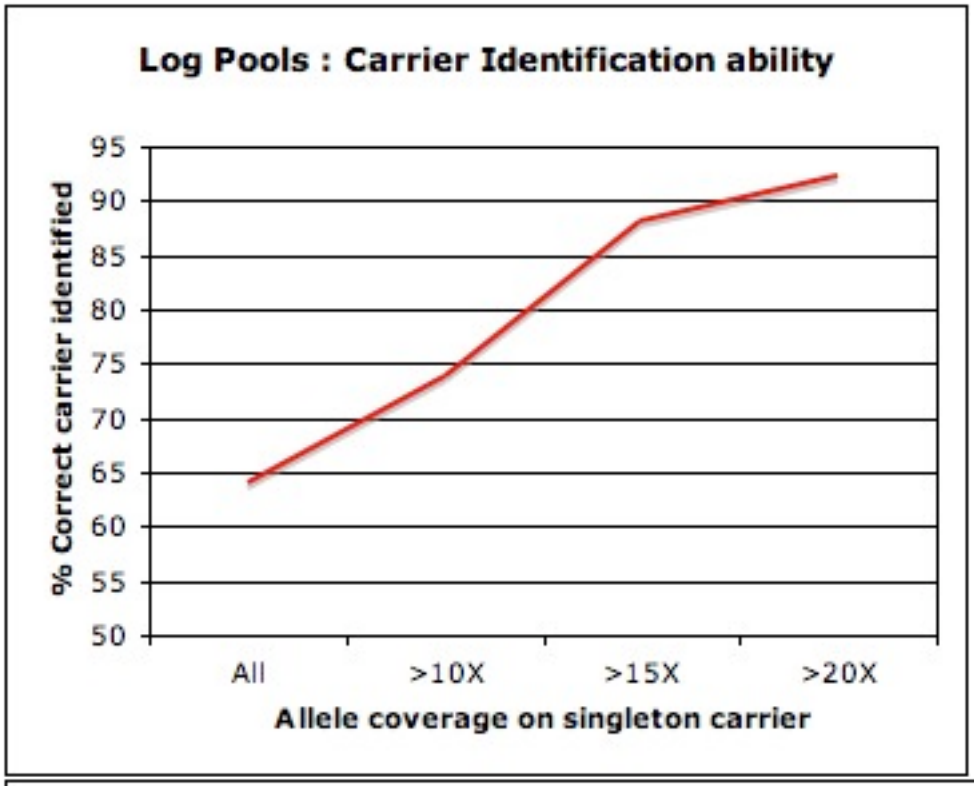
Log Pools:

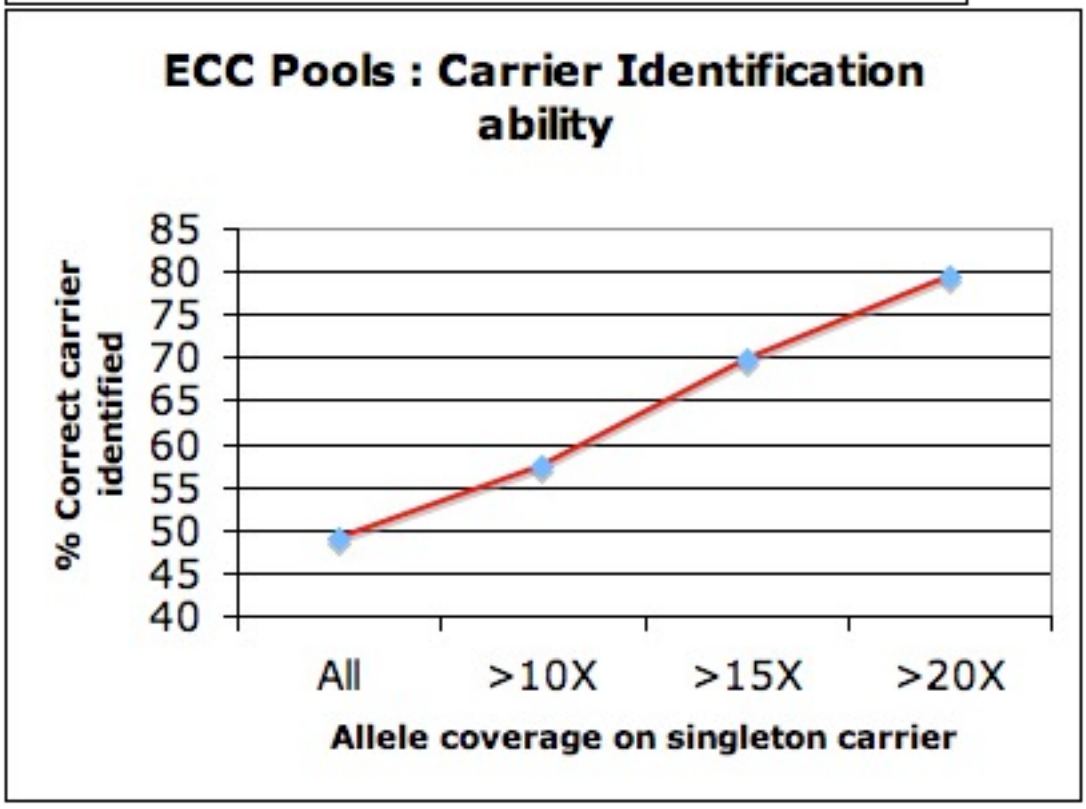
- 345 out of a total of 539 single carrier identifications were correct across all coverage profiles.
- 311 out of a total of 421 single carrier identifications were correct when the site had greater 10X allele coverage in the carrier individual's dataset.
- 266 out of a total of 302 single carrier identifications were correct when the site had greater than 15X allele coverage in the carrier individual's dataset.
- 206 out of a total of 223 single carrier identifications were correct when the site had greater than 20X allele coverage in the carrier individual's dataset.

ECC Pools:

- 783 out of a total of 1597 single carrier identifications were correct across all coverage profiles.
- 637 out of a total of 1109 single carrier identifications were correct when the site had greater 10X allele coverage in the carrier individual's dataset.
- 441 out of a total of 633 single carrier identifications were correct when the site had greater than 15X allele coverage in the carrier individual's dataset.
- 321 out of a total of 405 single carrier identifications were correct when the site had greater than 20X allele coverage in the carrier individual's dataset.

The result confirms our belief that ECC pools have a higher ability to identify singleton carriers. The graphs below chart the increase in fidelity of the call as coverage changes.


Log Pools : Carrier Identification ability
% Correct carrier identified
95
90
85
80
75
70
65
60
55
50
All
>10X
>15X
>20X
Allele coverage on singleton carrier


ECC Pools : Carrier Identification ability
% Correct carrier identified
85
80
75
70
65
60
55
50
45
40
All
>10X
>15X
>20X
Allele coverage on singleton carrier