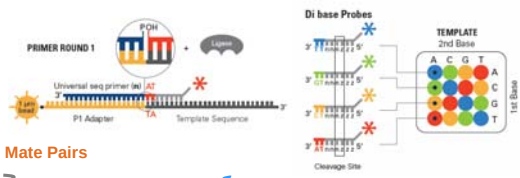
A high-resolution structural variation map of human germline genomes by next-generation, high-throughput paired end sequencing with the SOLiD System.

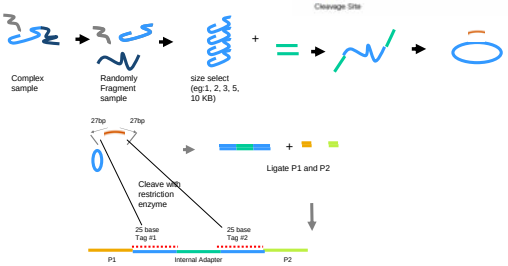
Heather Peckham#, Stephen F. McLauglin#, Swati Ranade#, C. Lee#, Y. Fu#, Zheng Zhang*, Fiona C.L. Hyland*, C. Clouser#, A. Antipova#, J. Manning#, C Hendrickson#, Gina Costa#, Francisco M. De La Vega*, and Kevin McKernan#. Applied Biosystems, *Foster City, CA, and #Beverly, MA.

MATERIALS AND METHODS

SOLiD System™ Chemistry



Mate Pairs



Sample

We sequenced NA18507, a Yoruba male HapMap sample.

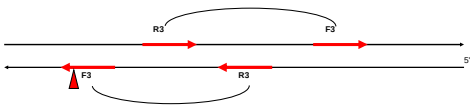
Runs

- 7 Fragment libraries, 50bp reads
  - Up to 6 Gbp mappable (usable) data per run
  - 10.5 Gbp - 7X sequence coverage
- 7 Paired end libraries, 25bp X 2 (6 libraries) & 35bp X 2 (1 library)
  - 17.6 Gbp - 5X sequence coverage
  - 136X physical coverage (includes insert size)
- 7 runs in total (2 slides/libraries per run), $60,000 reagent list price
- Total sequence coverage 12X
- Data deposited at the NCBI Short Read Archive Acc. No. 272

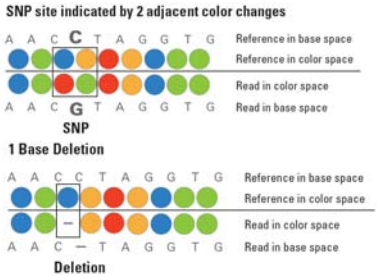
Mate Pairs: Detection of Large InDels

Orientation:	F3 and R3 on same strand
Ordering:	R3 is 5' of F3
Distance:	Determined by library

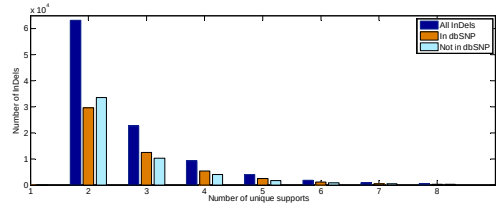
Mate Pairs: Detection of Small InDels



Dibase Chemistry enables accurate detection of SNPs, InDels



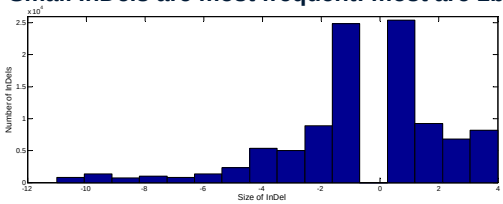
102,134 Small InDels detected (1 – 10 bp)



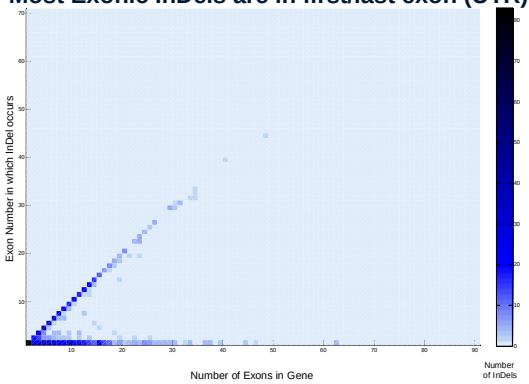
We detect 102,134 small InDels, 50.4% of which are seen in dbSNP. Each of these InDels has at least two independent reads with different start points confirming its existence (38% of them had 3 or more reads as evidence).

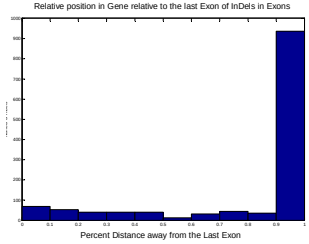
We detect deletions to 10bp, insertions to 4bp using mate pairs and re-mapping one tag allowing for an insertion or deletion.

Small InDels are most frequent: most are 1bp



Most Exonic InDels are in first/last exon (UTR)

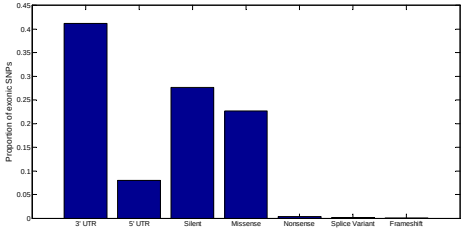




Of 1290 small InDels in exons, most are in first/last exon, consistent with being in 3' or 5' UTR.

- 210 are in the first exon
- 851 are in the last exon
- (82 in first and last)
- 311 are in an internal exon

Most SNPs are Intergenic. 54% of coding SNPs are silent, 45% missense, 0.6% nonsense



- 1.98% of SNPs are in exons (exons are under-represented).
- 36,654 of 335,836 exons (10.9%) contain at least one SNP .
- The vast majority of SNPs in and around genes are intronic.
- 81.4% of SNPs we detect are in dbSNP.
- 67.85% of heterozygous SNPs are transitions
  - 66.3% of novel heterozygous SNPs are transitions
  - 68.8% of known heterozygous SNPs are transitions.

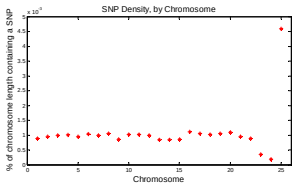
Damaging SNPs are overrepresented in genes for Olfaction and Immunity, and underrepresented in genes for transcription factors, ligases, growth factors, receptors

We annotate the functions of genes using the Panther ontology, and we annotate the damaging potential of non-synonymous SNPs (nsSNPs) using PolyPhen.

20.5% of nsSNPs are predicted to be possibly or probably damaging.

We discovered that **transcription factors, ligases, growth factors, receptors, and RNA helicases** are the molecular functions most under-represented for damaging mutations. No genes in any of these classes had a single damaging mutation. Further, we discovered that **GPCR genes involved in Olfaction**, and genes for **Immunity and defense** are the biological functions most highly over-represented for damaging mutations.

SNP density by chromosome



- Mitochondria have the highest SNP density
- Sex chromosomes have lowest SNP density
- SNP density on autosomes is ~ 1/1000 bp

CONCLUSIONS

Next generation sequencing has the potential to enable important applications in human genetics, including the detection of large and small InDels, of Inversions, and of homozygous and heterozygous SNPs.

Mate pairs of varies sizes facilitate discovery of structural variation including small InDels, with few runs. Dibase encoding facilitates accurate SNP detection at low coverage. Cost-effective whole genome sequencing is now feasible.