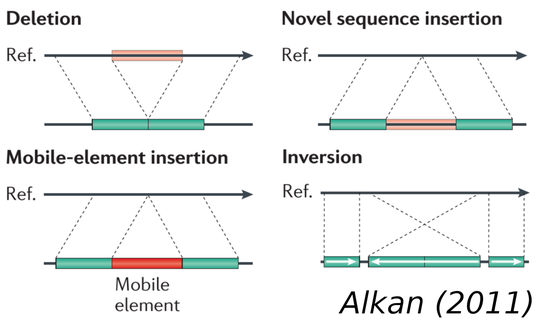
# Improving the reference with a diversity panel of sequence-resolved structural variation

Peter A. Audano[1,9], Arvis Sulovari[1,9], Tina A. Graves-Lindsay[2], Stuart Cantsilieris[1], Melanie Sorensen[1], AnneMarie E. Welch[1], Max L. Dougherty[1], Bradley J. Nelson[1], Ankeeta Shah[3], Susan K. Dutcher[2], Wesley C. Warren[2], Vincent Magrini[4,5], Sean D. McGrath[4], Yang I. Li[6,7], Richard K. Wilson[4,5], and Evan E. Eichler[1,8]

1. Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA; 2. McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA; 3. Committee on Genetics, Genomics, and Systems Biology, University of Chicago, Chicago, IL, USA; 4. Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, USA; 5. The Ohio State University College of Medicine, Columbus, OH, USA; 6. Section of Genetic Medicine, University of Chicago, Chicago, IL, USA; 7. Department of Human Genetics, University of Chicago, Chicago, IL, USA; 8. Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA; 9. These authors contributed equally to this work.

## Motivation

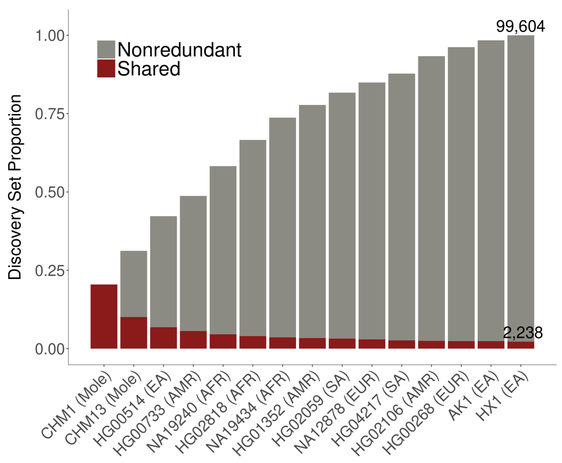


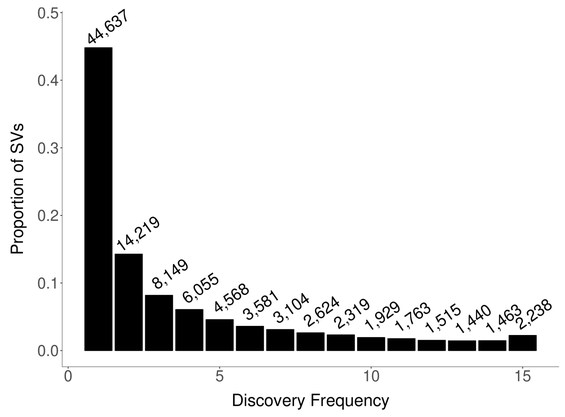

*Alkan (2011)*

Structural variants (SVs) are insertions, deletions, and inversions 50 bp or larger. Short-read sequence data is abundant, but it cannot identify most SVs or sequence-resolve them. Long reads (10-20 kbp, PacBio) are capable of resolving 90% of the genome, but a diversity panel of samples was not available until recently.

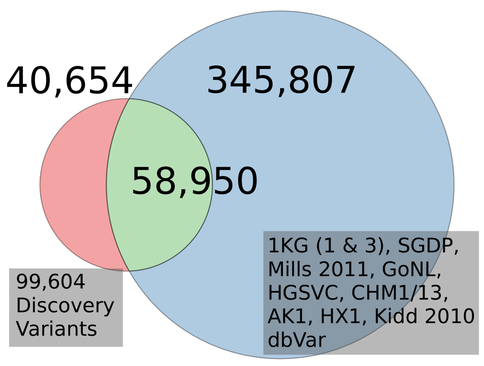
We sequenced 13 genomes with PacBio and obtained sequence-resolved SVs. With these data, we can better understand SV biology, correct the reference, and make a diverse set of variation accessible to short-read technology.

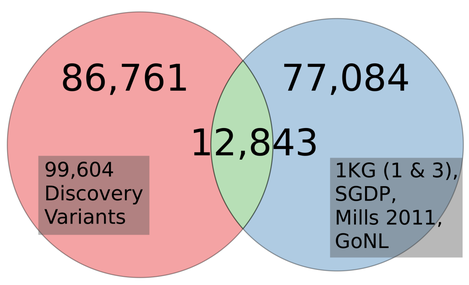
## Variant Discovery



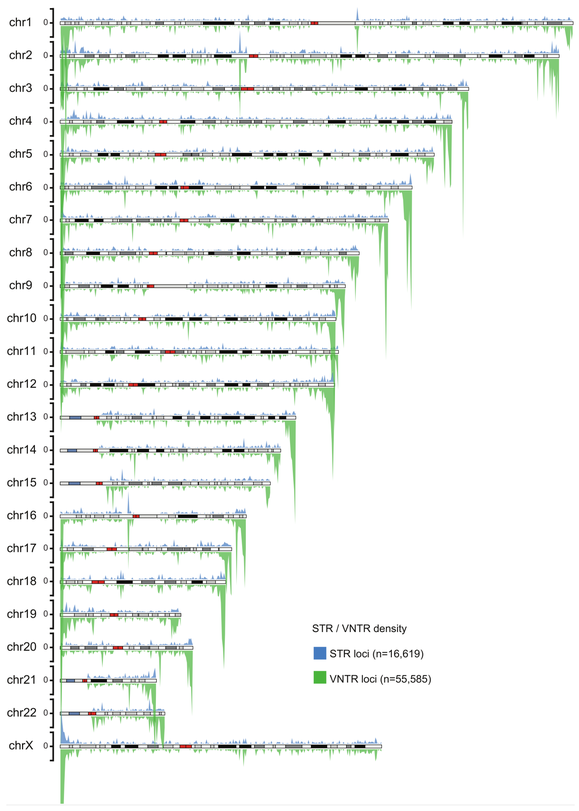




SVs were discovered[1] in our 13 genomes along with with AK1[2] and HX1[3], covering African, Asian, European, American, and S. Asian populations. In the nonredundant set of 99,604 SVs, 2,238 (1.6 Mbp) were shared among all 15 samples indicating errors or extreme minor alleles in GRCh38.




## STR and VNTR Enrichment

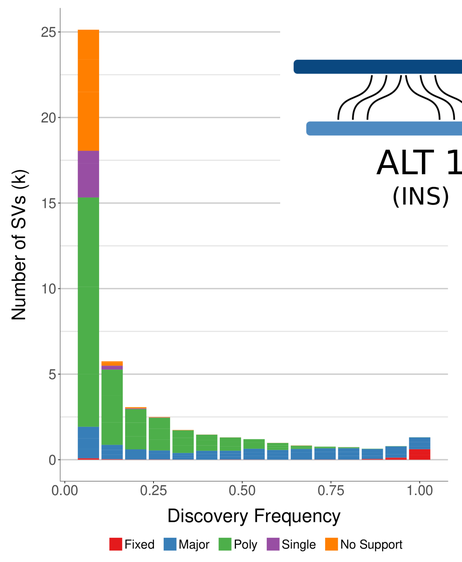


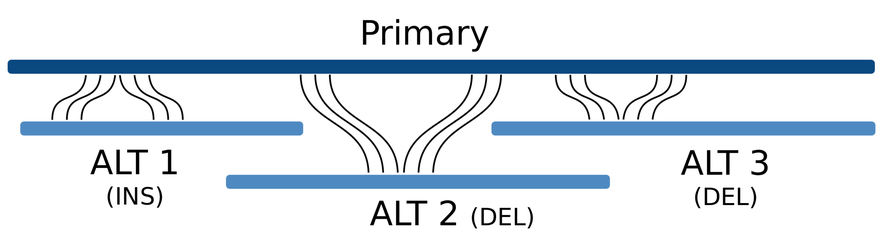

| Factor | Variance Relative | Variance Total | Test Statistic Corr. Coef. | Test Statistic p (F-test) |
|---|---|---|---|---|
| VNTR | | 43% | 0.66 | $<2.2\times10^{-16}$ |
| MR | 6% | | | |
| DSB | 5.70% | | | |
| MR x DNM | 4.40% | | | |
| MR X DSB | 3.90% | | | |
| STR | | 25% | 0.49 | $<2.2\times10^{-16}$ |
| MR | 6.60% | | | |
| DSB | 5.40% | | | |
| MR x DSB | 2.50% | | | |
| DSB x DNM | 1.60% | | | |
| No Repeat | | 24% | 0.49 | $<2.2\times10^{-16}$ |
| MR | 4.20% | | | |
| DSB | 3.40% | | | |
| DNM | 3% | | | |
| MR x DSB | 2.10% | | | |
| All SVs | | 36% | 0.60 | $<2.2\times10^{-16}$ |
| MR | 6.90% | | | |
| DSB | 5.20% | | | |
| MR X DSB | 3.50% | | | |
| DNM | 2.80% | | | |

SVs are enriched in the last 5 Mbp of chromosome arms. This effect is driven by VNTRs, which exhibit a 4.8-fold enrichment (Wilcoxon $p = 2.9 \times 10^{-9}$). Male meiotic recombination (MR), double-strand breaks (DSB), and *de novo* mutations (DNM) were most significantly correlated suggesting potential mechanisms of VNTR formation.

*Credit: Arvis Sulovari*

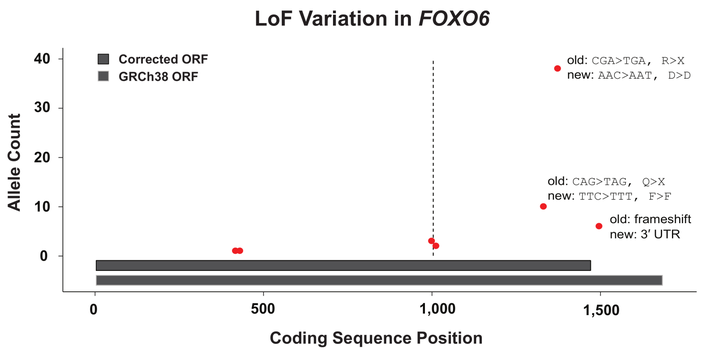
## Building a Patched Reference




By adding sequence-resolved SVs to the reference in alternate contigs, we can improve read mapping over SVs and accurately genotype known variants in short-read samples with a machine learning model. With our patched reference, we called variants in SV insertions and found 2,228 SNPs and indels (GQ 20+) per sample that would be missed by standard pipelines.

## Gene and Regulatory Impact




200 bp missing in *FOXO6* shifts the RefSeq annotation out of frame. gnomAD[4] contains three common mutations past the insertion (dashed line), but when corrected, two become synonymous variants and one falls into the UTR.

*Credit: Max Dougherty*

| Class | CDS N | CDS % | UTR N | UTR % | NC Regulatory N | NC Regulatory % | Intron/2kbp Flank N | Intron/2kbp Flank % | All Gene/Reg N | All Gene/Reg % |
|---|---|---|---|---|---|---|---|---|---|---|
| Shared | 5 | 0.22% | 6 | 0.27% | 160 | 7.15% | 1,111 | 49.64% | 1,180 | 52.73% |
| Major | 81 | 0.62% | 41 | 0.31% | 873 | 6.69% | 6,306 | 48.31% | 6,712 | 51.42% |
| Polymorphic | 326 | 0.82% | 161 | 0.41% | 2,410 | 6.07% | 18,969 | 47.81% | 20,262 | 51.07% |
| Singleton | 429 | 0.96% | 286 | 0.64% | 2,838 | 6.36% | 20,653 | 46.27% | 22,232 | 49.81% |
| All | 841 | 0.84% | 494 | 0.50% | 6,281 | 6.31% | 47,039 | 47.23% | 50,386 | 50.59% |

CDS: Coding regions. UTR: Untranslated region. NC Regulatory: H3K4Me1, H3K4Me3, H3K27Ac, DNase hypersensitivity

## Conclusions and Future Work

- A population-level view of SVs supports corrections across the genome
- SVs impact genes, regulatory elements, and their annotations
- SVs are non-randomly distributed over the genome
- VNTRs correlate with double-strand breaks and male meitic recombination
- Sequence-resolution enables genotyping with short reads

We plan to:

- Sequence 50 additional genomes (PacBio, 10X)
- Apply phasing (Phased-SV[5]) to improve SVs and contigs
- Build a pan-genome reference (vg[6])

## References

1. Huddleston, J. *et. al.* (2017). Genome Research
2. Seo, J. *et. al.* (2016). Nature
3. Shi, L. *et. al.* (2016). Nature Communications
4. Lek, M. *et. al.* (2016). Nature
5. Human Genome Structural Variant Consortium. (2017). BioRxiv. https://doi.org/https://doi.org/10.1101/193144
6. Garrison, E. *et. al.* (2018). Nature Biotechnology