bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity

Qingyu Wang[1,2], Cooduvalli S. Shashikant[1,3], Naomi S. Altman[1,2], Santhosh Girirajan[1,4,5*]

[1]Bioinformatics and Genomics Program, The Huck Institutes of the Life Sciences; [2]Department of Statistics; [3]Department of Molecular and Developmental Biology; [4]Department of Biochemistry and Molecular Biology; and [5]Department of Anthropology, The Pennsylvania State University, University Park, PA16802

Correspondence to:

Santhosh Girirajan, MBBS, PhD

205A Life Sciences Building

The Pennsylvania State University

University Park, PA16802

E-mail: sxg47@psu.edu

Phone: 814-865-067

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ABSTRACT

Whole Exome Sequencing (WES) is a powerful clinical diagnostic tool for discovering the genetic basis of many diseases. A major shortcoming of WES is uneven coverage of sequence reads over the exome targets contributing to many low coverage regions, which hinders accurate variant calling. In this study, we devised two novel metrics, Cohort Coverage Sparseness (CCS) and Unevenness (U_E) Scores for a detailed assessment of the distribution of coverage of sequence reads. Employing these metrics we revealed non-uniformity of coverage in the WES data generated by three different platforms. This non-uniformity of coverage is both local (coverage of a given exon across different platforms) and global (coverage of all exons across the genome in the given platform). The low coverage regions encompassing functionally important genes were often associated with high GC content, repeat elements and segmental duplications. Some of the problems associated with WES are due to extrapolation of technologies primarily designed for Whole Genome Sequencing platforms. Further refinements in these technologies have potential to enhance the clinical applications of WES platforms.

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

INTRODUCTION

Whole Exome Sequencing (WES) is a high throughput genomic technology that sequences coding regions of the genome selectively captured by target enrichment strategies[1]. Target enrichment is achieved by oligonucleotide probes that selectively hybridize and capture the entire coding region of the genome referred to as the exome[1-3]. Since the exome represents approximately two percent of the genome, WES technology provides high coverage at a lower cost and in a shorter time compared to Whole Genome Sequencing (WGS) technology[4]. From its first successful application in discovering the candidate gene associated with Miller syndrome[5], WES has been used to study a number of Mendelian[6] and complex disorders[7-10]. WES is used in the 1000 Genomes and the NHLBI GO exome sequencing projects, to catalog variants in the population and to identify rare variants associated with diseases[11-14]. Since 2011, WES has also been routinely offered as a diagnostic tool in clinical genetics laboratories[15-17]. A recent study reported that in a large cohort of patients referred by a physician to receive genetic testing, 25% of the patients received a genetic diagnosis, including diseases such as neurodevelopmental disorders, cancer, cardiovascular disease, and immune-related diseases[18].

Several target enrichment strategies to capture exomes are available, including the widely used Agilent SureSelect Human All Exon capture kit, Roche NimbleGen SeqCap EZ Exome capture system, and the Illumina TruSeq Exome Enrichment kit. While the basic sample preparation protocols are similar among these platforms, major differences lie in the design of the oligonucleotide probes, including selection of target genomic regions, sequence features and lengths of probes, and the exome capture mechanisms. This may lead to some differences in genes captured on each chromosome by the different platforms. In spite of its extensive use, the analysis of WES data still presents considerable challenges. There are significant concerns

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

regarding unevenness of sequence read coverage, which affects downstream analysis. For example, even in samples with high average read depths (>75X), some regions are captured poorly (with coverage as low as 10X), potentially resulting in missed variant calls[19]. Studies examining the overall quality of WES data have focused on comparing the performance of a single DNA sample or a small number (n≤6) of samples in different capture technologies[20-22]. While these studies have focused on the GC content and overall coverage differences between different platforms, the intra-platform variation in sequence coverage, characteristics of the low-coverage regions, and variation of coverage across the exome have not been evaluated. We undertook a comprehensive assessment of sequence coverage in the human exome, and examined the variance in read depth between samples and across the exome. We evaluated the sequence content and characteristics of the genomic regions contributing to systematic biases in exome sequencing using WES data from a total of 176 individuals obtained from three different platforms. Our study provides quantitative metrics for systematic analysis of different parameters that could potentially impact WES analysis and confirms the association between low coverage regions and occurrence of duplicated sequences and high GC content.

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RESULTS

To assess the coverage distribution of reads, we selected 176 out of 212 exome sequence samples obtained from NimbleGen, Agilent and Illumina TruSeq platforms, with an average target coverage of at least 75X. Even after normalizing for the library size, the coverage at specific positions in the exons varied among different samples for each of the three platforms tested. For example, the average read depth of exon 16 of *TP53BP2* in two samples sequenced with NimbleGen at a similar average coverage (92X) was 48X and 92X, respectively. This inconsistency in coverage distribution resulted in a range of 10X-500X read depth at several regions in the exome when multiple samples were run on the same platform. Such regions of highly variable read coverage mapping within disease-associated genes can affect the accuracy of current algorithms for variant calling in genetic studies. To characterize the distribution of sequence reads along the exome, we developed two metrics, Cohort Coverage Sparseness (CCS) and the Unevenness (U_E) scores. The CCS score provides an assessment of coverage of all exons across the genome (global) in the given platform, while the U_E score provides an assessment of coverage of a given exon (local) across different platforms.

Coverage deficiencies determined by CCS Score

Read coverage of a position in the genome is considered deficient if the number of reads mapped to that position is less than 10[23]. The CCS score is defined as the percentage of low coverage (<10X) bases within a given exon in multiple WES samples. The CCS score is estimated by first calculating the read depth for each base position in a given exon and then determining the median percentage of samples with low coverage at that region (see Methods). The resulting CCS score may vary between 0 and 1, with high CCS scores indicating low sequence coverage.

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We plotted CCS scores for all genes against the chromosome positions in a modified Manhattan plot (**Figure 1**). As shown in the figure, a majority of genes (>88%) were clustered in the low CCS score region (<0.2), indicating good coverage in all three platforms (**Figure 1A-C**). The remaining genes with CCS scores >0.2, which contain low coverage regions, were scattered throughout the plot. In those low coverage genes, CCS score >0.5 indicates nearly half of the regions have less than 10X read depth. The distribution of CCS scores is skewed to the right when plotted as a histogram (**Figure S1**), consistent with the pattern shown in **Figure 1**. Further, the data generated from Illumina TruSeq have the lowest percentage of low coverage genes (~7%), compared with data generated by the capture kits from NimbleGen (~9%) and Agilent (~11%), as shown in **Table 1**. The differences in these percentages could be due to differences in the design of probes that target the exome in these platforms[20]. **Table S1** lists all low coverage genes identified in the three WES platforms compared to a WGS dataset.

The low coverage regions (CCS>0.2) varied significantly across all chromosomes in the three platforms analyzed (**Table S2**; chi-square test p-value $<2.2\times10^{-16}$). Chromosomes 6 and 19 have a higher proportion of low coverage genes compared to other chromosomes. A low coverage gene cluster on chromosome 6 (cytobands 6p21.33 and 6p21.32) corresponded to the genes encoding human leukocyte antigen, which are known to be polymorphic and have alleles showing high sequence identity[24,25]. Similarly, chromosome 19 is known to carry a high proportion of tandem gene families, repeat sequences, and segmental duplications (SD)[26]. These sequence features potentially affect accurate mapping of reads, leading to low coverage regions.

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Coverage non-uniformity determined by U_E Score

The U_E score provides a measure of non-uniformity calculated after smoothing the coverage distribution curves and identifying peaks and troughs along the curve (**Figure 2**, see Methods). It is calculated based on the number and structural features (height, width, base) of the coverage peaks. The U_E score increases with an increase in the number and relative height of peaks within a given exon. For regions with uniformly distributed coverage, the U_E score is 1; for regions with uneven coverage, the U_E score is greater than 1. To illustrate how U_E score varies among different platforms, we chose the last coding exon of *ZNF484*, which had high (CCS scores < 0.01) but variable (U_E score 72.2-141.8) coverage across the different platforms, indicating inconsistent coverage (**Figure 3**).

We also observed a positive association between exon length and the U_E score in all three platforms as shown in the scatter plot (**Figure 4**). The Pearson correlation coefficient for all platforms was ≥0.7 (NimbleGen, 0.80, Agilent, 0.71; TruSeq, 0.70). The U_E score was significantly different for longer exons (>400 bp) among the three platforms tested (Friedman test, p-value $<2.2\times10^{-16}$). A post hoc analysis[27] also showed that the U_E score was highest for Illumina, followed by NimbleGen and Agilent (**Figure S2**). When the coverage distribution of the neighboring exons was examined, we found inconsistencies in the rank order of read coverage among samples tested on the same platform (**Figure S3**). Since most of the current methods for calling copy number variations (CNVs) are based on detecting continuous depletion or enrichment after normalizing for coverage of adjacent exons[28-30], this lack of consistency of rank orders between closely occurring exons could significantly affect CNV calling from exon read depths.

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Low coverage regions are enriched within repeat elements

Since chromosomes 6 and 19 contain repeat elements and clustered gene families that accounted for specific regions showing low coverage, we checked if these sequence features globally correspond with the occurrence of low coverage. In one study, SD regions from the Agilent WES dataset and the Segmental Duplication Database (http://humanparalogy.gs.washington.edu) were cross-referenced with the low coverage genes obtained from our analysis, and examined for the percentage of reads that mapped unambiguously to off-target regions. The percentage of off-target reads with multiple hits in SD regions (50%) was about double that of those found in unique regions (26%). These results suggest that the mapping method may also contribute to the missing coverage.

Because SD regions share similar features with repeat elements, we next tested if low coverage regions were associated with underlying repeat sequences. We first examined the 358 bp exon 1 of *MAST4*, which contains a 22 bp low complexity repeat element predicted by RepeatMasker (http://www.repeatmasker.org)[31]. As shown in **Figure 5**, the coverage was high (30X) for sequence base positions 1-200 in samples with high average coverage. The read depth at each nucleotide in this region remains high. In contrast, the coverage falls dramatically (to <10X) between base positions 200-358, with corresponding low read depth at each nucleotide in this region. The low coverage region is greater than a typical exome capture probe size (50-150 bp), indicating that influences of a repeat element can extend beyond its location.

To test for a global association between the occurrence of repeat elements and low coverage, we examined all the troughs (from U_E score calculations) in coverage in the 176 samples studied. Repeat elements predicted by RepeatMasker coincided with extremely low troughs (median coverage <10X) for a majority of the exons (60% to 69% of exons tested on the

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

three platforms), especially for exons >200 bp in size. For example, as shown in **Figure S4**, multiple repeat elements occur in the extremely low trough region (positions 1 to 400 bp) in the last coding exon of *CASZ1*. However, we also find that several troughs contain repeat elements that are not annotated by RepeatMasker when default parameters are used (**Figure S5**). This suggests that coverage troughs are highly associated with repeat sequences.

Negative association between GC content and sequence coverage

Since GC content is one of the major factors contributing to low coverage in WES data[20,32], we investigated whether there is an association between GC content and CCS score. We used a density curve plot and a modified density plot, which visualizes the distribution of GC content of genes with the corresponding CCS scores (**Figure 6**, **Figure S6**). Based on GC content, we were able to clearly distinguish exons with CCS score <0.2 from exons with CCS score >0.2. Thus, high GC content correlated with high CCS scores, and therefore with lower coverage regions. Exons with good sequence coverage (CCS<0.2) were clustered in the intermediate GC content regions (between 30 -70%). In contrast, a higher density of low coverage exons (CCS>0.2) was observed in regions with relatively high GC content (>70%). No good coverage exons were present in regions with less than 20% or more than 80% GC content. However, some poorly covered exons (CCS>0.8) were found in the intermediate GC content regions (<70% and > 30%), suggesting that there may be other factors contributing to the low coverage within these regions.

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Low coverage regions contain functionally relevant genes

To examine whether low coverage regions had functional significance, we conducted Functional Disease Ontology Annotations (FunDO)[33] of 827 low coverage genes that were common to all three platforms (**Figure 7A**). Result showed enrichment of genes implicated in cancer, diabetes mellitus, and heart failure (**Figure 7B**). We also examined the coverage of genes that American College of Medical Genetics and Genomics (ACMG) recommends for pathogenic variant discovery and clinical reporting[34]. Of the 56 genes examined, seven genes, including *BRCA2*, *KCNH2*, *KCNQ1*, *SDHD*, *TNN3*, *VHL*, and *WT1*, mapped within low coverage regions in one or more samples (with average coverage >75X) (**Figure S7**). These results suggest that low coverage regions within functionally important genes could affect variant discovery and subsequent clinical diagnosis.

Low coverage is more of an issue for WES than WGS based platforms

We compared the exon coverage between the WES and WGS datasets to test whether mapping issues were common to both platforms. We examined the CCS scores for all genes (**Figure 1D**), and the U_E scores (**Figure 4)** and GC contents (**Figure 6D**) for all exons. For WGS datasets with an average coverage of >30X (Table S1), only 30 genes with high CCS scores were observed in the modified Manhattan plot (**Figure 1D**). In contrast, for WES platforms with an average coverage of >75X, over 1,000 genes with high CCS scores were observed (**Figure 1A-C**). Similarly, the U_E score for all exons in the WGS analysis was significantly lower compared to the WES analysis (Friedman's test, $p<2.2\times10^{-16}$). As shown in **Figure 4**, the U_E score increased slightly with increased size of the exon. To examine how GC content affects coverage in the WGS dataset, we generated probability density curves for GC content for different gene groups

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(**Figure 6**). In all the three WES platforms, we found a dramatic shift in the density plots for high coverage genes compared to low coverage genes based on GC content. In comparison, for the WGS platform, the shift in GC content between the exons with high and low CCS scores was minimal. Thus, GC content has less influence on sequence coverage in WGS compared to WES analysis. These results suggest that problems of low coverage are specific to WES platforms. The methods developed for WGS analysis require further modifications before application to WES platforms.

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DISCUSSION

WES is a powerful clinical diagnostic tool for identifying disease-associated variants in patients[35,36]. Since most known Mendelian disorders are associated with mutations in the coding regions, focusing on sequencing exomes rather than whole genomes is efficient in terms of time, expense and coverage[4]. Recent studies have successfully used WES technology to identify variants that strongly correlate with disease phenotypes[17,18,37,38]. However, high-resolution examination of different WES datasets shows uneven coverage along the length of exons, suggesting possible problems in variant calling. This affects identification of de novo variations that may be clinically important. In this study, we systematically examined different parameters that could potentially impact WES analysis and identified key issues associated with sequence architecture contributing to the low coverage.

We analyzed WES data captured by three major platforms: NimbleGen CapSeq V2, Agilent SureSelect V2 and Illumina TruSeq. All three platforms are based on similar target enrichment protocols and cover >95% of RefSeq[39] coding region with >90% low CCS scores. These platforms differ from one another in the layout of probes. Most importantly, while NimbleGen uses overlapping probes, Agilent uses tiling probes, and Illumina uses gapped probes[20]. This difference in probe design significantly contributes towards inconsistency in the coverage, creating systematic biases and preventing the combining of datasets from different platforms for SNV and CNV detection. This is a major concern, especially with large cohort studies involving multiple centers using different platforms. Even in the data generated by a single platform, the coverage distribution varied among different exons. Several factors including size of the exon, GC content, presence or absence of repeat elements, and segmental duplications affect the coverage. The uniformity of coverage distribution decreases for longer

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

exons. For a given exon, the pattern of coverage varied among different platforms even for genes with high coverage (**Figure 3**), making it difficult to normalize the background coverage for CNV calling. As shown in our study, coverage in regions with extremes of GC content (<30%; >70%) was low reflecting poor capture efficiency. In contrast, no such correlation was observed for regions with moderate GC content (30-70%). The coverage was also affected by the presence of repeat elements. Even simple repeat elements as small as 22 bp may contribute to low coverage. Those platforms, which use RepeatMasker in probe design, are likely to miss exonic regions containing repetitive elements.

Several studies comparing technologies have found WGS to be superior to WES[40-42]. The parameters affecting coverage including GC content, uniformity of the coverage, and read depths were examined. Our results are consistent with these findings. Although WGS data has lower average coverage, the coverage depth along exons is more uniformly distributed. WGS data have fewer sparse regions, which may contribute to lower numbers of false negative variant calls. However, with more than 100,000 exomes sequenced to date, WES has become the major genetic tool in several diagnostic centers[36,43]. A thorough understanding of the limitations of each of the WES platforms is thus important. Modifications in the design of the targeted sequence capture technology and improvements in mapping algorithms are essential for accurate calling of variants and filling the gaps in heritability estimates of genetic disease.

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MATERIALS AND METHODS

Datasets and Pre-processing

Datasets: WES data generated from NimbleGen SeqCap, Agilent SureSelect, and Illumina TruSeq were obtained for analysis from the dbGaP (the database of Genotypes and Phenotype)[44] and SRA (Sequence Read Archive) databases[45]. WGS sample data were obtained from the 1000 Genomes project server in BAM format. Detailed information for each dataset is listed in **Table S3**.

Mapping of reads: All SRA files of exome samples downloaded from different databases were first converted to FASTQ files using the SRA Toolkit[45]. Then, raw sequence reads were mapped to the reference genome using Bowtie2 version 2.1.0, with default parameters[46]. We used the hg19/GRCh37 assembly of the human genome as the reference sequence throughout the analysis, and SAMtools version 1.2 was used to sort reads and remove PCR duplicates[47]. The WGS dataset was downloaded as mapped reads in BAM format. The mapping was done at the Broad Institute using the Burrows Wheeler Aligner (BWA).

Exome and gene sets annotation: Overlapping exons of the same gene were merged to create a consensus exome annotation file from the RefSeq database using an in-house pipeline. Target regions obtained from three different capture kits were mapped and intersected with the consensus exome annotation file. These target regions were then sorted by genomic coordinates on chromosomes and defined as target exon regions. For WGS data analysis, all exons in the consensus exome annotation file were used as target exon regions. Only autosomal chromosomes 1 to 22 were included in this analysis.

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Coverage calculation: BEDTools[48] was used to calculate the single base pair coverage of all BAM files at all positions covered in the exome annotation file.

Metrics to Evaluate Low Coverage Regions

CCS score: The CCS score for all exons and genes was calculated with the ExomeCQA program. The CCS score provides a measure of percentage of base pairs in a given region for which coverage is less than 10 reads in multiple samples (see **Figure S9**). The CCS score is calculated by the formula:

$$\text{CCS} = median_N\left(\frac{\#(c_i<10)}{L_R}\right)$$

where, N is the total number of samples; # is the count of the genomic positions covered by read depth coverage less than 10x. c_i is the read depth coverage at genomic position i. L_R is the length of the region of interest.

U_E of coverage: The U_E metric was introduced to measure the non-uniformity of the coverage over targeted regions, and is calculated using ExomeCQA. The score was developed based on detecting the structures, peaks and troughs in coverage distribution curves. For a specific targeted region, the LOWESS (locally weighted scatterplot smoothing)[49] method was applied to the median coverage of all samples at each genomic position along the whole region to obtain a smoothed curve. Using the LOWESS function, we implemented a locally weighted regression of the read depth coverage (Y-axis) for each base position in a continuous genomic region (X-axis). We used an empirical span of 0.03 for the smoothing parameter. Peaks and troughs of the read depth coverage were identified by a Hill Climbing local optimization algorithm, which was

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

implemented by scanning all base positions and identifying the positions with increasing or decreasing neighbors. The height and width of the peaks were then used to calculate the unevenness score by the following formula:

$$U_E = \sum_{1}^{N} \frac{H/B}{W/L_R}$$

where U_E is unevenness score, N is number of peaks, H is height of the peak, B is base of the peak, and W is width of the peak.

Program ExomeCQA: ExomeCQA, was written to calculate different metrics of cohort exome sequencing data from coverage files input. This program was written in C++ and can be downloaded at http://exomecqa.sourceforge.net.

Statistical Methods

All the statistical tests were conducted in R. The Manhattan plots of CCS scores in all chromosomes were generated by the "GWASTool"[50] package. The density plot of GC content and low coverage regions was generated by the package "GenePlotter"[51]. Both packages are in Bioconductor[52]. The post hoc analysis for Friedman's test was conducted by the package "Coin"[53].

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES

1 Hodges, E. *et al.* Genome-wide in situ exon capture for selective resequencing. *Nat Genet* **39**, 1522-1527, doi:10.1038/ng.2007.42 (2007).

2 Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* **106**, 19096-19101, doi:10.1073/pnas.0910672106 (2009).

3 Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272-276, doi:10.1038/nature08250 (2009).

4 Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. Disease gene identification strategies for exome sequencing. *European journal of human genetics : EJHG* **20**, 490-497, doi:10.1038/ejhg.2011.258 (2012).

5 Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**, 30-35, doi:10.1038/ng.499 (2010).

6 Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am J Hum Genet* **97**, 199-215, doi:10.1016/j.ajhg.2015.06.009 (2015).

7 Jeste, S. S. & Geschwind, D. H. Disentangling the heterogeneity of autism spectrum disorder through genetic findings. *Nat Rev Neurol* **10**, 74-81, doi:10.1038/nrneurol.2013.278 (2014).

8 Wong, L. P. *et al.* Deep whole-genome sequencing of 100 southeast Asian Malays. *Am J Hum Genet* **92**, 52-66, doi:10.1016/j.ajhg.2012.12.005 (2013).

9 de Bruin, C. & Dauber, A. Insights from exome sequencing for endocrine disorders. *Nat Rev Endocrinol* **11**, 455-464, doi:10.1038/nrendo.2015.72 (2015).

10 O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-250, doi:10.1038/nature10989 (2012).

11 Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).

12 Regalado, E. S. *et al.* Exome sequencing identifies SMAD3 mutations as a cause of familial thoracic aortic aneurysm and dissection with intracranial and other arterial aneurysms. *Circ Res* **109**, 680-686, doi:10.1161/CIRCRESAHA.111.248161 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

13 Emond, M. J. *et al.* Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis. *Nat Genet* **44**, 886-889, doi:10.1038/ng.2344 (2012).

14 Boileau, C. *et al.* TGFB2 mutations cause familial thoracic aortic aneurysms and dissections associated with mild systemic features of Marfan syndrome. *Nat Genet* **44**, 916-921, doi:10.1038/ng.2348 (2012).

15 Pierson, T. M. *et al.* Whole-exome sequencing identifies homozygous AFG3L2 mutations in a spastic ataxia-neuropathy syndrome linked to mitochondrial m-AAA proteases. *PLoS Genet* **7**, e1002325, doi:10.1371/journal.pgen.1002325 (2011).

16 Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-241, doi:10.1038/nature10945 (2012).

17 Yang, Y. *et al.* Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med* **369**, 1502-1511, doi:10.1056/NEJMoa1306555 (2013).

18 Yang, Y. *et al.* Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* **312**, 1870-1879, doi:10.1001/jama.2014.14601 (2014).

19 Hoischen, A., Krumm, N. & Eichler, E. E. Prioritization of neurodevelopmental disease genes by discovery of new mutations. *Nat Neurosci* **17**, 764-772, doi:10.1038/nn.3703 (2014).

20 Clark, M. J. *et al.* Performance comparison of exome DNA sequencing technologies. *Nat Biotechnol* **29**, 908-914, doi:10.1038/nbt.1975 (2011).

21 Sulonen, A. M. *et al.* Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol* **12**, R94, doi:10.1186/gb-2011-12-9-r94 (2011).

22 Asan *et al.* Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biol* **12**, R95, doi:10.1186/gb-2011-12-9-r95 (2011).

23 Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nature reviews. Genetics* **15**, 121-132, doi:10.1038/nrg3642 (2014).

24 Dawkins, R. *et al.* Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease. *Immunol Rev* **167**, 275-304 (1999).

25 Williams, T. M. Human leukocyte antigen gene polymorphism and the histocompatibility laboratory. *J Mol Diagn* **3**, 98-104, doi:10.1016/S1525-1578(10)60658-7 (2001).

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

26 Grimwood, J. *et al.* The DNA sequence and biology of human chromosome 19. *Nature* **428**, 529-535, doi:10.1038/nature02399 (2004).

27 Gerald, K. B. Nonparametric statistical methods. *Nurse Anesth* **2**, 93-95 (1991).

28 Kadalayil, L. *et al.* Exome sequence read depth methods for identifying copy number changes. *Briefings in bioinformatics* **16**, 380-392, doi:10.1093/bib/bbu027 (2015).

29 Fromer, M. *et al.* Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet* **91**, 597-607, doi:10.1016/j.ajhg.2012.08.005 (2012).

30 Krumm, N. *et al.* Copy number variation detection and genotyping from exome sequence data. *Genome Res* **22**, 1525-1532, doi:10.1101/gr.138115.112 (2012).

31 Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **Chapter 4**, Unit 4 10, doi:10.1002/0471250953.bi0410s05 (2004).

32 Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* **12**, R18, doi:10.1186/gb-2011-12-2-r18 (2011).

33 Osborne, J. D. *et al.* Annotating the human genome with Disease Ontology. *BMC Genomics* **10 Suppl 1**, S6, doi:10.1186/1471-2164-10-S1-S6 (2009).

34 Green, R. C. *et al.* ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med* **15**, 565-574, doi:10.1038/gim.2013.73 (2013).

35 Helbig, K. L. *et al.* Diagnostic exome sequencing provides a molecular diagnosis for a significant proportion of patients with epilepsy. *Genet Med*, doi:10.1038/gim.2015.186 (2016).

36 Stark, Z. *et al.* A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders. *Genet Med*, doi:10.1038/gim.2016.1 (2016).

37 Talkowski, M. E., Minikel, E. V. & Gusella, J. F. Autism spectrum disorder genetics: diverse genes with diverse clinical outcomes. *Harv Rev Psychiatry* **22**, 65-75, doi:10.1097/HRP.0000000000000002 (2014).

38 Berg, J. S. Genome-scale sequencing in clinical care: establishing molecular diagnoses and measuring value. *JAMA* **312**, 1865-1867, doi:10.1001/jama.2014.14665 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

39 Pruitt, K. D., Tatusova, T., Klimke, W. & Maglott, D. R. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res* **37**, D32-36, doi:10.1093/nar/gkn721 (2009).

40 Meynert, A. M., Ansari, M., FitzPatrick, D. R. & Taylor, M. S. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics* **15**, 247, doi:10.1186/1471-2105-15-247 (2014).

41 Lelieveld, S. H., Spielmann, M., Mundlos, S., Veltman, J. A. & Gilissen, C. Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. *Hum Mutat* **36**, 815-822, doi:10.1002/humu.22813 (2015).

42 Belkadi, A. *et al.* Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A* **112**, 5473-5478, doi:10.1073/pnas.1418631112 (2015).

43 Monroe, G. R. *et al*. Effectiveness of whole-exome sequencing and costs of the traditional diagnostic trajectory in children with intellectual disability. *Genet Med*, doi:10.1038/gim.2015.200 (2016).

44 Mailman, M. D. *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* **39**, 1181-1186, doi:10.1038/ng1007-1181 (2007).

45 Leinonen, R., Sugawara, H., Shumway, M. & International Nucleotide Sequence Database, C. The sequence read archive. *Nucleic Acids Res* **39**, D19-21, doi:10.1093/nar/gkq1019 (2011).

46 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

47 Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993, doi:10.1093/bioinformatics/btr509 (2011).

48 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

49 Cleveland, W. S. Weighted Regression and Smoothing Scatterplots. *J Am Stat Assoc* **74**, 829-836 (1979).

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

50 Gogarten, S. M. *et al*. GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics* **28**, 3329-3331, doi:10.1093/bioinformatics/bts610 (2012).

51 Gentleman, R. C. & Biocore. geneplotter: Graphics related functions for Bioconductor. (R package version 1.48.0.).

52 Gentleman, R. C. *et al*. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80, doi:10.1186/gb-2004-5-10-r80 (2004).

53 Hothorn, T., Hornik, K., Wiel, M. A. v. d. & Zeileis, A. Implementing a class of permutation tests: The coin package. *J Stat Softw* **28**, 1-23 (2008).

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgements: We thank DebMalya Nandy, Matthew Jensen and Emily Huber from the Girirajan lab for critical reading of the manuscript. This work was supported by a Basil O'Connor Award from the March of Dimes Foundation (#5-FY14-66), a NARSAD Young Investigator Grant from the Brain and Behavior Research Foundation, and resources from the Huck Institutes of the Life Sciences (SG), and The Pennsylvania State University Experiment Station grant AES 4586 (CCS).

We acknowledge the NIH data repository, the contributing investigators, and the associated primary funding organizations for providing access to the datasets used in this study. We thank NIH Genomic Study Datasets, Christopher A. Walsh, MD, PhD. Children's Hospital Boston, Boston, MA, USA (PI of the dataset), and their funding resource: U54 HG003067. National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA; the Department of Health Sciences, University of Milano–Bicocca, Monza, Italy (for SRP028277); Liver Cancer Institute, Zhongshan Hospital, Fudan University, Grant: ID 81272725, "Identification of driving mutations and singling pathway alterations in intrahepatic cholangiocarcinoma", National Natural Science Foundation of China (for SRP025150); McGill University, Cancer Research Society, Hungarian Scientific Research Fund (OTKA) contract T04639, Canadian Institutes of Health Research grant 102684, and National Research and Development Fund (NKFP) (for SRP032767). Funding support for the dataset in "Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations" (phs000482.v1.p1) was provided by the Simons Foundation Autism Research Initiative (SFARI 137578 & 191889), NIH HD065285 and the Howard Hughes Medical Institute. Data was originally reported by O'Roak et al. 2012, PMCID: PMC3350576. We are grateful to all of the families at the participating SFARI Simplex Collection (SSC) sites, as well as the principal

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, E. Hanson, D. Grice, A. Klin, R. Kochel, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, E. Wijsman).

Author Contributions: Q.W. and S.G. conceived the problem. Q.W., S.G., N.S.A., C.S.S. contributed to the design of computational approaches. Q.W. performed all the analysis, developed algorithms and the Rescuer program. Q.W. and C.S.S. wrote the manuscript N.S.A. and SG contributed to the revisions.

Competing Financial Interests: The authors declare no competing financial interests.

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. Low coverage genes with high CCS scores in three different datasets

Platforms	Genes*	Genes with CCS > 0.2 (%)	Genes with CCS > 0.5 (%)
NimbleGen	19,956	1,796 (9%)	418 (2%)
Agilent	18,409	2,025 (11%)	374 (2%)
Illumina TruSeq	17,886	1,252 (7%)	228 (1.3%)

*Gene sets in different platforms were defined as described in the Methods section

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

FIGURE LEGENDS

Figure 1: CCS scores of targeted RefSeq genes along the whole chromosome in WES and WGS datasets. The CCS values are plotted along the length of each chromosome in a modified Manhattan Plot for WES datasets obtained from (**A**) NimbleGen, (**B**) Agilent, (**C**) Illumina TruSeq, and (**D**) WGS dataset from 1000 genome project.

Figure 2: Characterizing the coverage distribution with the Unevenness (U_E) score. (**A**) The read depth coverage distribution from multiple samples is plotted against the exon length. (**B**) The smoothed median coverage plotted against the exon length, obtained by first calculating median coverage for each position and then using LOWESS smoothing. Peaks and troughs were then identified by using a local optimization algorithm. Arrows indicate peaks identified in the curve: B, base, W, width and H, height of the peak., L_R, length of the region analyzed

Figure. 3: Base coverage distribution along the length of the last coding exon of gene *ZNF484* from WES datasets obtained from (**A**) NimbleGen, (**B**) Agilent, and (**C**) Illumina TruSeq.

Figure 4: Scatterplot of Unevenness (U_E) scores against exon size in WES and WGS datasets.

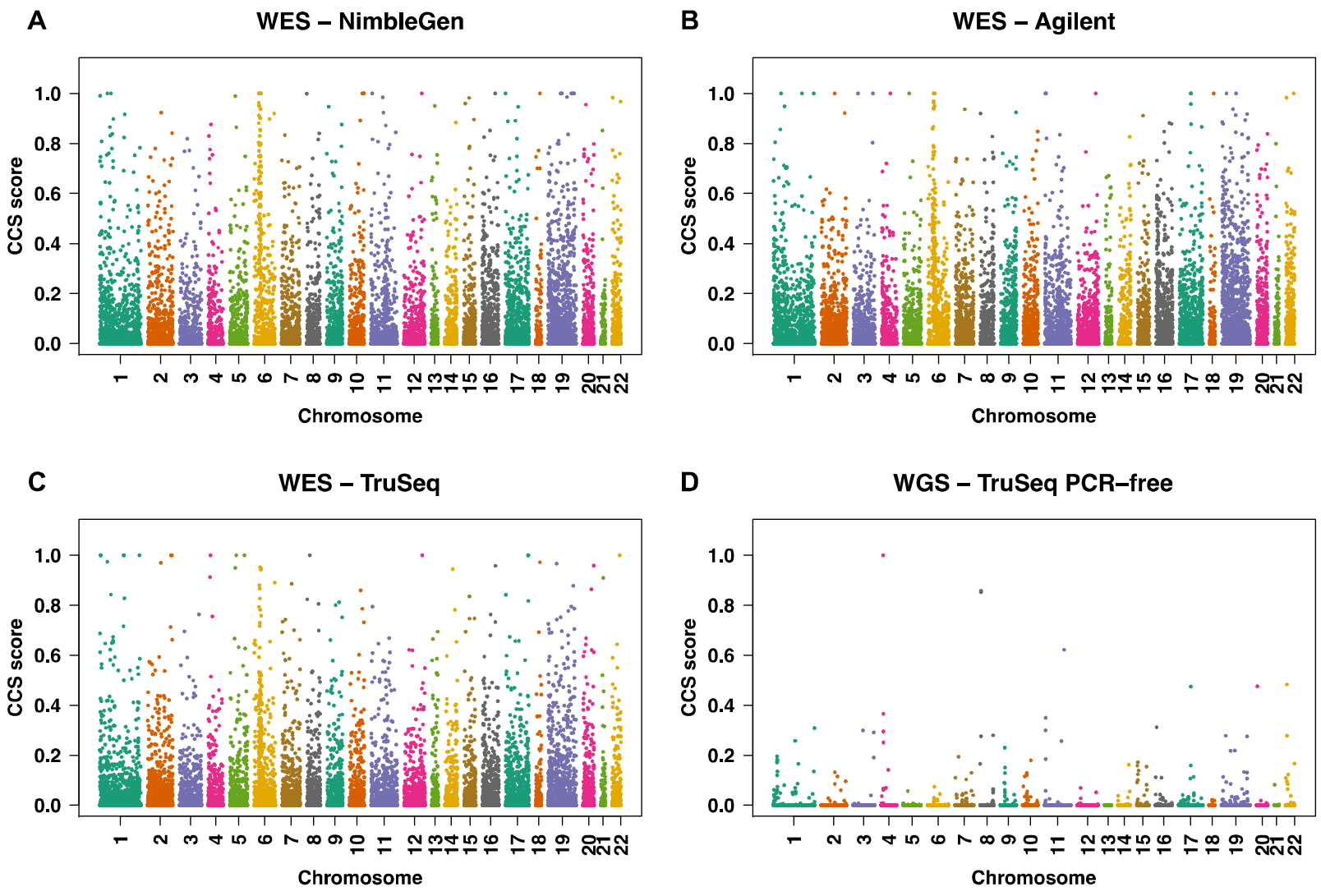
Figure. 5: Concurrence of repeat elements and coverage sparseness. (**A**) Base coverage distribution along the length of the first coding exon of MST4. WES samples with different average coverage ranging from 75X to 200X are shown in different colors. Arrow indicates the point at which coverage falls sharply. (**B**) UCSC browser screen shot of MST4 genomic region, black bar indicates the position of the repeat element.

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure. 6: The probability density curves showing GC content in sets of genes with different coverage. The distribution of GC content of all genes (black), high coverage genes with CCS score < 0.2 (green), low coverage genes with CCS score > 0.2 (blue) are represented.

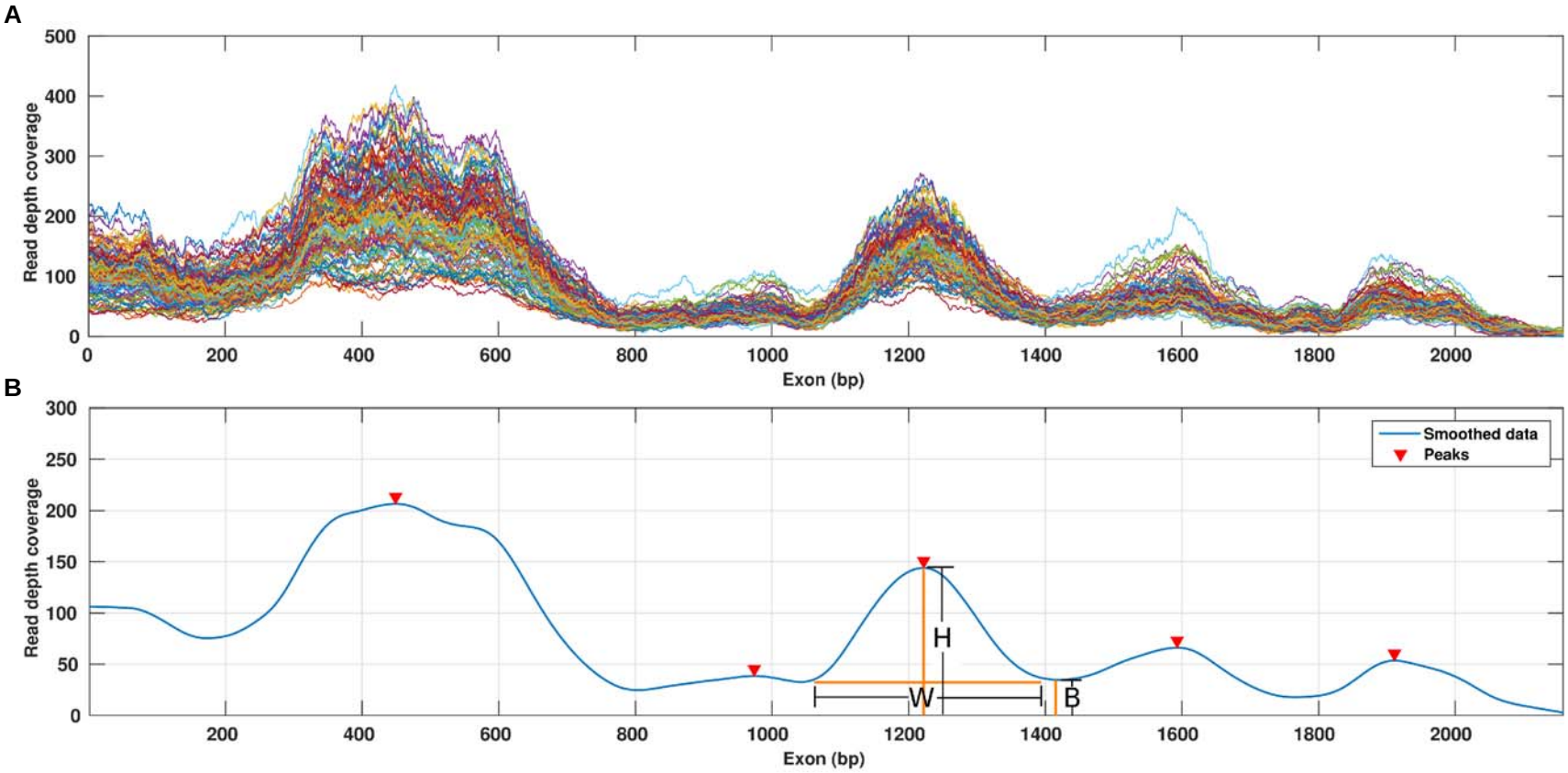
Figure. 7: Genes with low coverage in three different datasets. (**A**) Venn diagram showing number of low coverage genes (CCS score > 0.2) across three different platforms. There are 827 genes with low coverage in common across all platforms. (**B**) Network diagram showing disease ontology analysis of the 827 low-coverage genes showing associations with cancer, leukemia, diabetes mellitus, heart failure, and mucocutaneous lymph node syndrome.

Figure 1



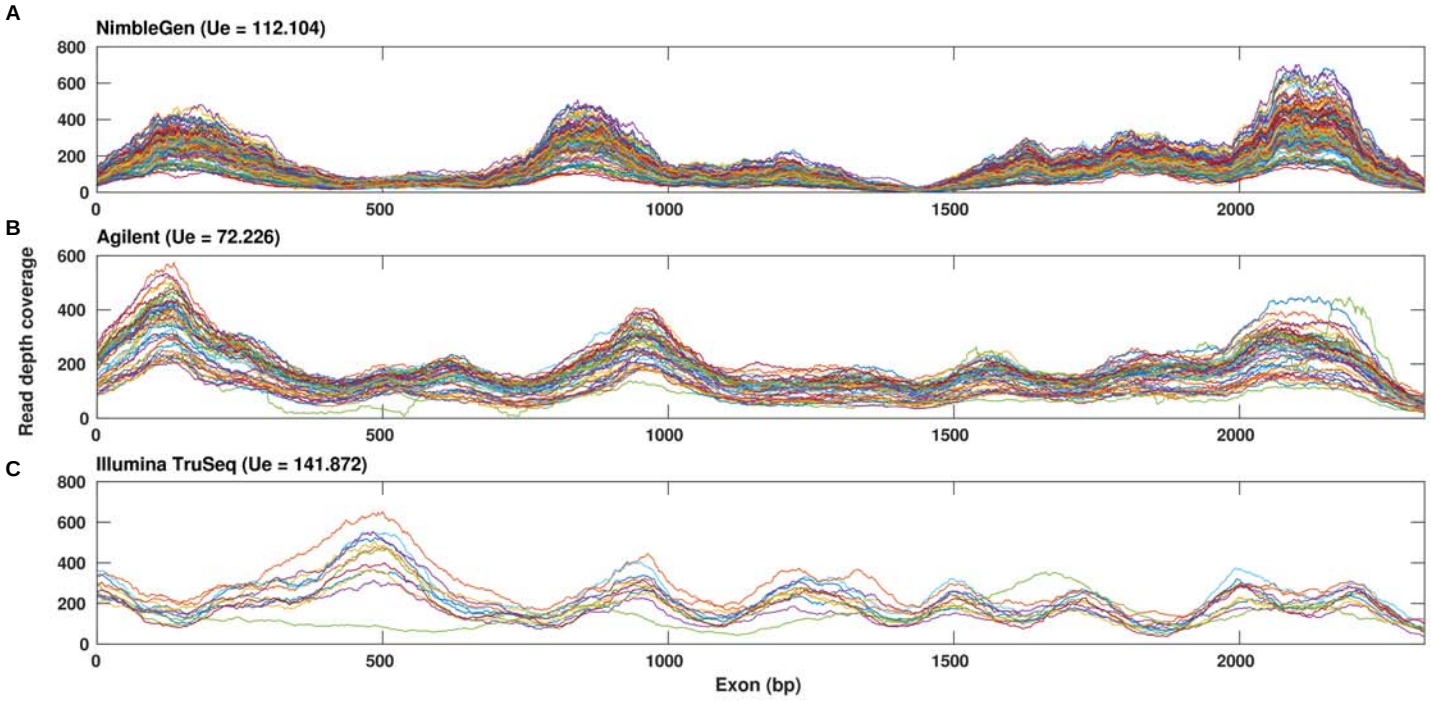
bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2



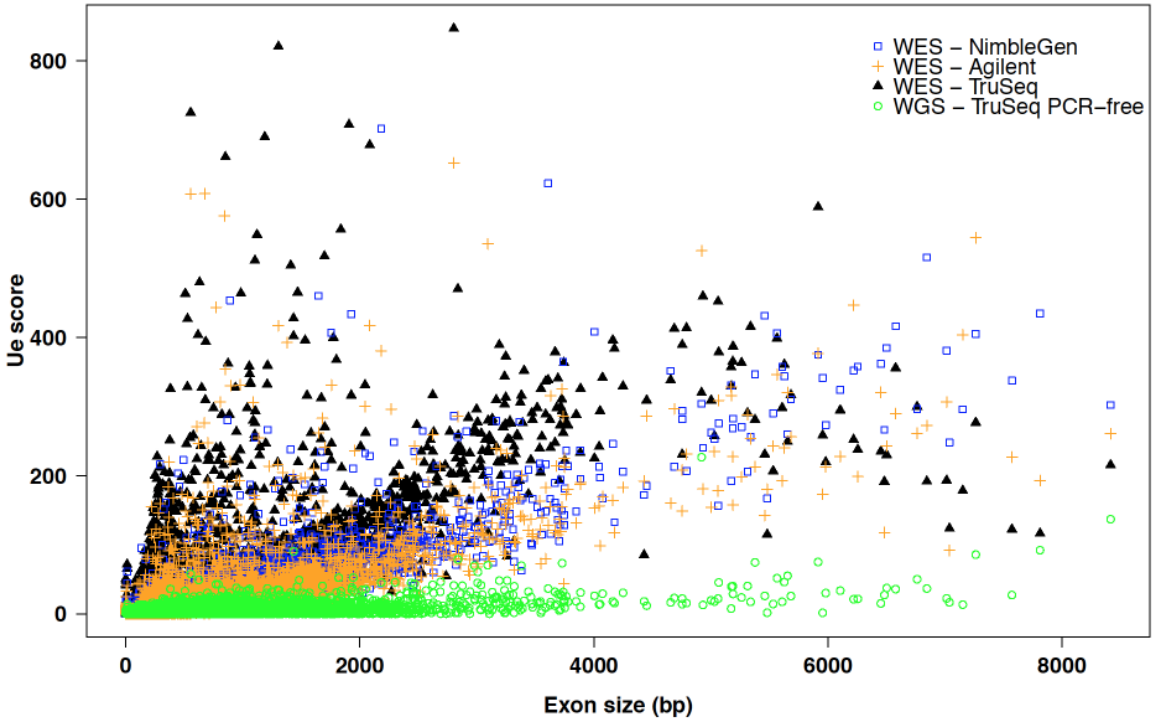
bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3



bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

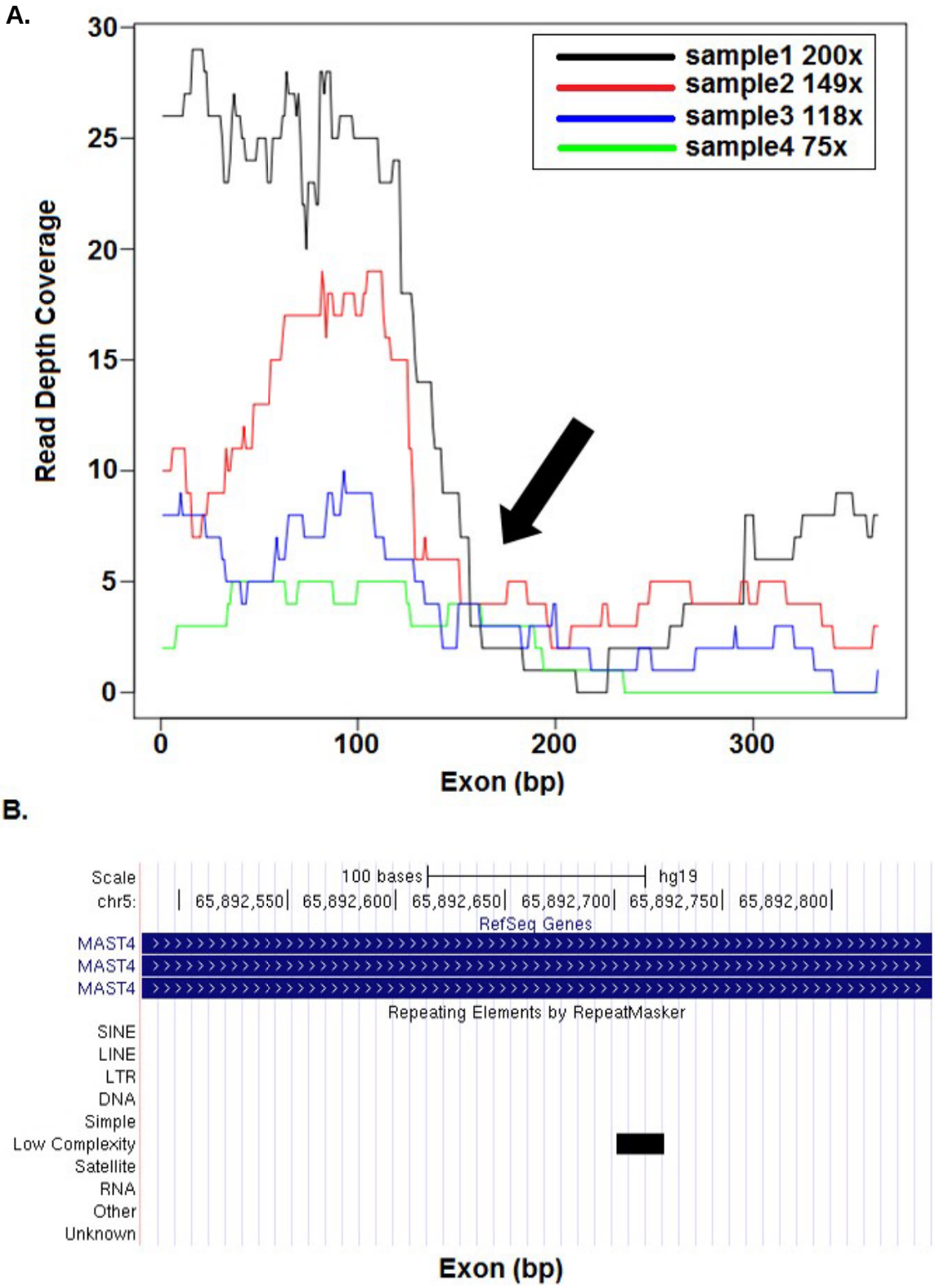
Figure 4



bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 5



bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 6

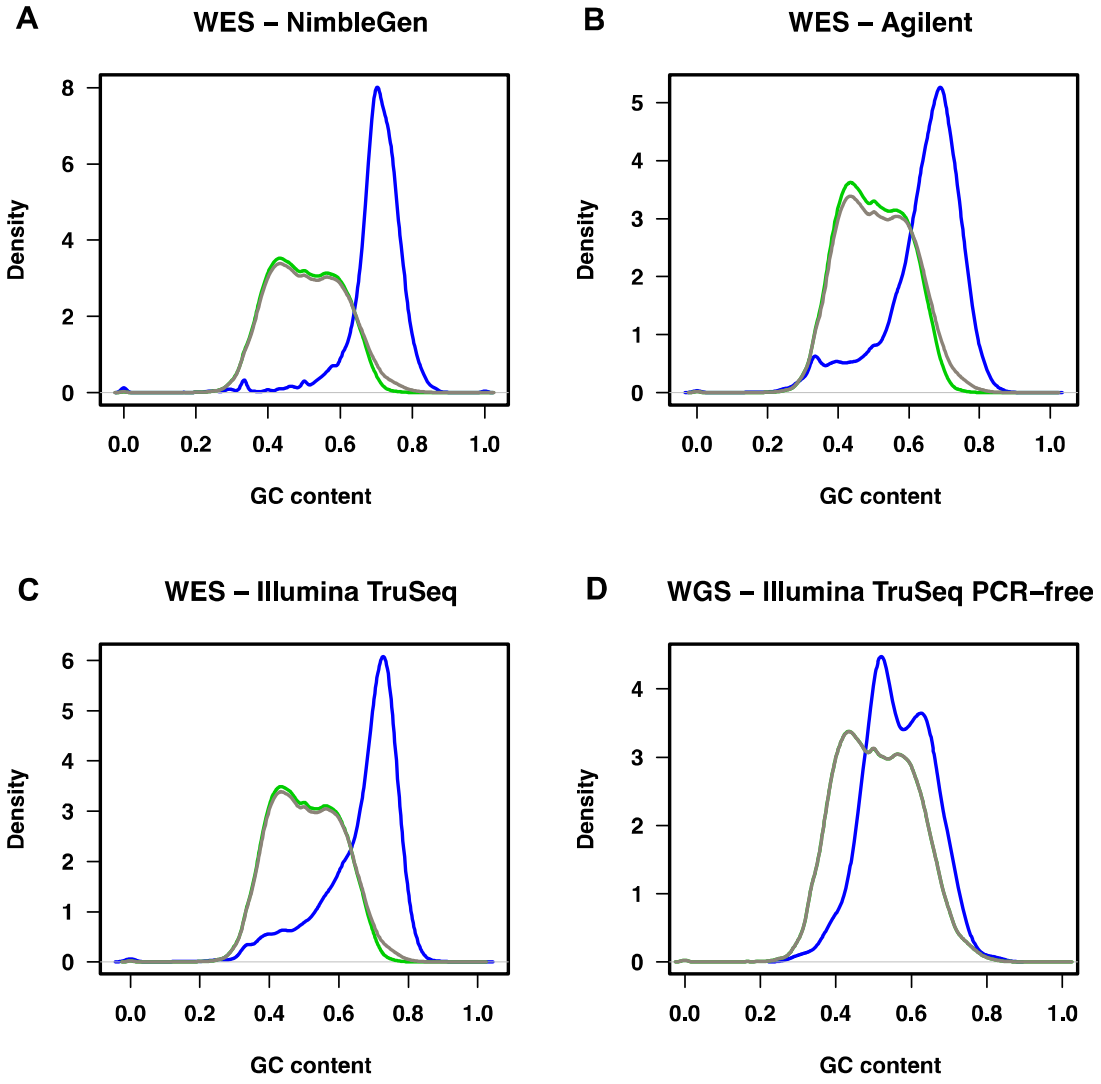
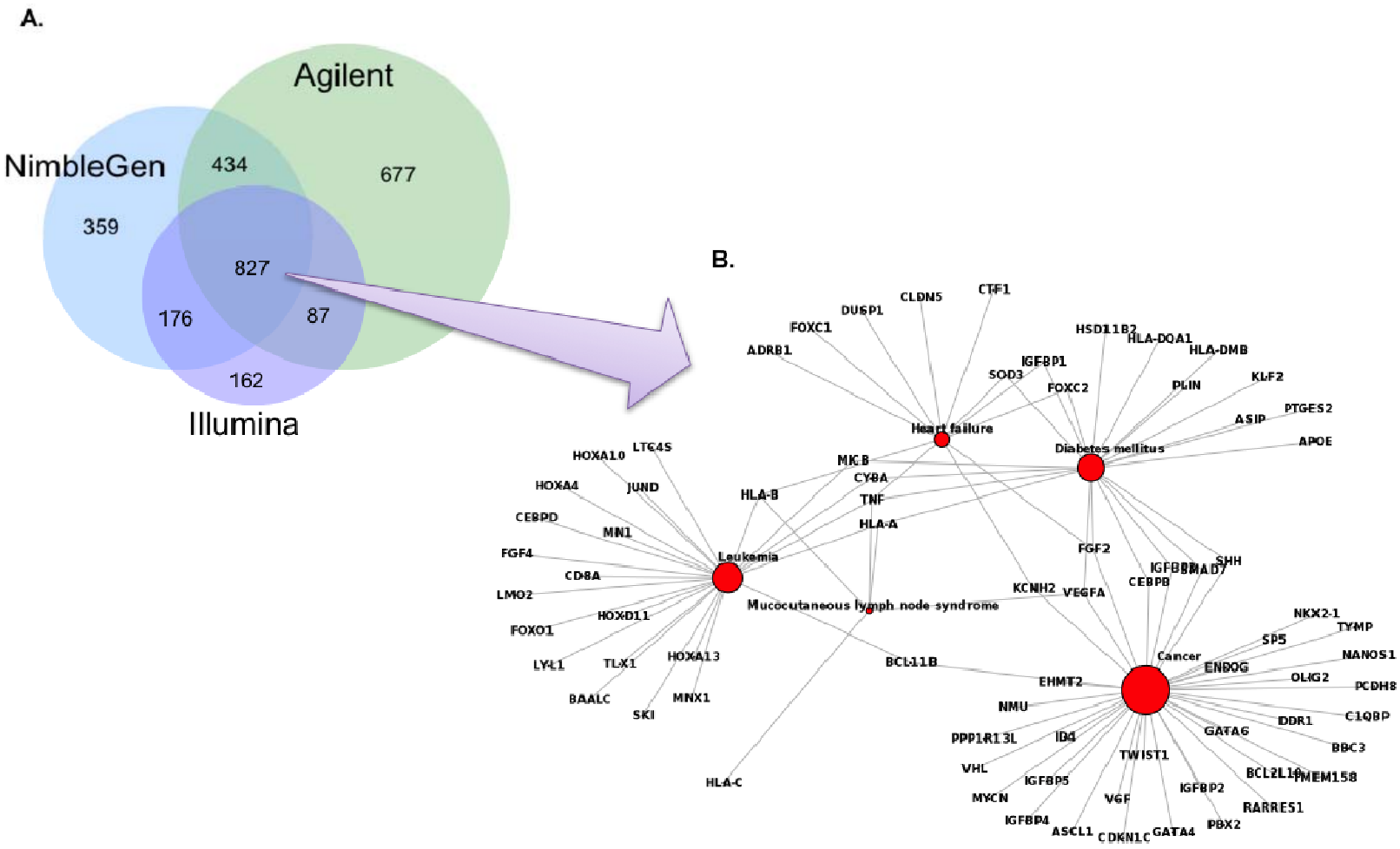


Figure 7



bioRxiv preprint doi: https://doi.org/10.1101/051888; this version posted May 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.